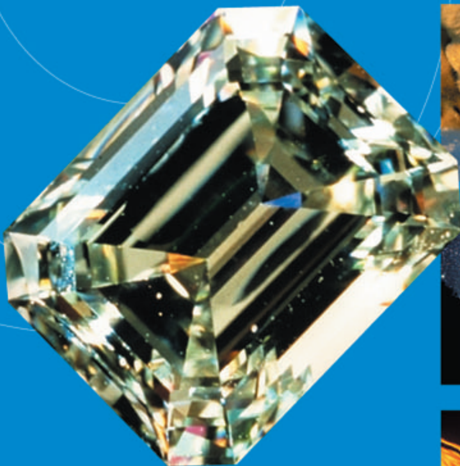
SCIENCE OF EVERYDAY THINGS

VOLUME 4: REAL-LIFE EARTH SCIENCE



EDITED BY NEIL SCHLAGER
WRITTEN BY JUDSON KNIGHT

SCIENCE OF EVERYDAY THINGS

VOLUME 2: REAL-LIFE PHYSICS

EDITED BY NEIL SCHLAGER
WRITTEN BY JUDSON KNIGHT

A SCHLAGER INFORMATION GROUP BOOK

GALE GROUP
THOMSON LEARNING

Detroit • New York • San Diego • San Francisco
Boston • New Haven, Conn. • Waterville, Maine
London • Munich

SCIENCE OF EVERYDAY THINGS

SCIENCE OF EVERYDAY THINGS

SCIENCE OF EVERYDAY THINGS

VOLUME 4: REAL-LIFE EARTH SCIENCE

EDITED BY NEIL SCHLAGER
WRITTEN BY JUDSON KNIGHT

A SCHLAGER INFORMATION GROUP BOOK

GALE®

THOMSON
GALE™

Detroit • New York • San Diego • San Francisco • Cleveland • New Haven, Conn. • Waterville, Maine • London • Munich

THOMSON
GALE

Science of Everyday Things
Volume 4: Real-Life Earth Science

A Schlager Information Group Book
Neil Schlager, Editor
Written by Judson Knight

Project Editor
Kimberley A. McGrath

Editorial
Mark Springer

Permissions
Kim Davis

Imaging and Multimedia
Robert Duncan, Leitha Etheridge-Sims, Mary K. Grimes, Lezlie Light, Dan Newell, David G. Oblender, Robyn V. Young

Product Design
Michelle DiMercurio, Michael Logusz

Manufacturing
Evi Seoud, Rhonda Williams

For more information contact
The Gale Group, Inc.
27500 Drake Rd.
Farmington Hills, MI 48331-3535
Or you can visit our Internet site at
http://www.gale.com

Cover photographs for volumes 1-4 reproduced by permission of: *VOLUME 1:* Dave Bartruff/Corbis (neon in Las Vegas, NV), Sergio Dorantes/Corbis (Coca-Cola bottles), Reuters NewMedia Inc./Corbis (Nissan Hypermini), Reuters NewMedia Inc./Corbis (PVC dress on runway); *VOLUME 2:* PhotoDisc 2001 (skate boarder), The Purcell Team/Corbis (roller coaster), Bob Rowan/Progressive Image/Corbis (grocery store scanner), Tim Wright/Corbis (crash test); *VOLUME 3:* BSIP/V & L /Photo Researchers (geneticist/scientist), Eye of Science/Photo Researchers (dustmite), Gallo Images/Corbis (African honeybee), Stuart Westmorland/Corbis (starfish); *VOLUME 4:* John Shaw/Photo Researchers (polar bear), F. Zullo/Photo Researchers (star gazer), V. Fleming/Photo Researchers (emerald-cut diamond), B. Edmaier/Photo Researchers (Mt. Etna lava flow).

LIBRARY OF CONGRESS CATALOG-IN-PUBLICATION DATA

Knight, Judson.
Science of everyday things / written by Judson Knight, Neil Schlager, editor.
p. cm.
Includes bibliographical references and indexes.
Contents: v. 1. Real-life chemistry – v. 2 Real-life physics.
SBN 0-7876-5631-3 (set : hardcover) – ISBN 0-7876-5632-1 (v. 1) – ISBN 0-7876-5633-X (v. 2)
1. Science–Popular works. I. Schlager, Neil, 1966-II. Title.

Q162.K678 2001
500–dc21 2001050121

ISBN 0-7876-5631-3 (set), 0-7876-5632-1 (vol. 1), 0-7876-5633-X (vol. 2), 0-7876-5634-8 (vol. 3), 0-7876-5635-6 (vol. 4)

Printed in the United States of America
10 9 8 7 6 5 4 3 2 1

Contents

INTRODUCTION

Overview of the Series

Welcome to *Science of Everyday Things*. Our aim is to explain how scientific phenomena can be understood by observing common, real-world events. From luminescence to echolocation to buoyancy, the series will illustrate the chief principles that underlay these phenomena and explore their application in everyday life. To encourage cross-disciplinary study, the entries will draw on applications from a wide variety of fields and endeavors.

Science of Everyday Things initially comprises four volumes:

Volume 1: *Real-Life Chemistry*
Volume 2: *Real-Life Physics*
Volume 3: *Real-Life Biology*
Volume 4: *Real-Life Earth Science*

Future supplements to the series will expand coverage of these four areas and explore new areas, such as mathematics.

Arrangement of Real-Life Earth Science

This volume contains 40 entries, each covering a different scientific phenomenon or principle. The entries are grouped together under common categories, with the categories arranged, in general, from the most basic to the most complex. Readers searching for a specific topic should consult the table of contents or the general subject index.

Within each entry, readers will find the following rubrics:

- **Concept:** Defines the scientific principle or theory around which the entry is focused.
- **How It Works:** Explains the principle or theory in straightforward, step-by-step language.
- **Real-Life Applications:** Describes how the phenomenon can be seen in everyday life.
- **Where to Learn More:** Includes books, articles, and Internet sites that contain further information about the topic.

In addition, each entry includes a "Key Terms" section that defines important concepts discussed in the text. Finally, each volume includes many illustrations and photographs throughout.

Included in this volume, readers will find (in addition to the volume-specific general subject index), a cumulative general index, as well as a cumulative index of "everyday things." This latter index allows users to search the text of the series for specific everyday applications of the concepts.

About the Editor, Author, and Advisory Board

Neil Schlager and Judson Knight would like to thank the members of the advisory board for their assistance with this volume. The advisors were instrumental in defining the list of topics, and reviewed each entry in the volume for scientific accuracy and reading level. The advisors include university-level academics as well as high school teachers; their names and affiliations are listed elsewhere in the volume.

Neil Schlager is the president of Schlager Information Group Inc., an editorial services company. Among his publications are *When*

Technology Fails (Gale, 1994); *How Products Are Made* (Gale, 1994); the *St. James Press Gay and Lesbian Almanac* (St. James Press, 1998); *Best Literature By and About Blacks* (Gale, 2000); *Contemporary Novelists*, 7th ed. (St. James Press, 2000); *Science and Its Times* (7 vols., Gale, 2000-2001); and *Science in Dispute* (Gale, 2002). His publications have won numerous awards, including three RUSA awards from the American Library Association, two Reference Books Bulletin/Booklist Editors' Choice awards, two New York Public Library Outstanding Reference awards, and a *CHOICE* award for best academic book.

Judson Knight is a freelance writer, and author of numerous books on subjects ranging from science to history to music. His work on science includes *Science, Technology, and Society, 2000 B.C.-A.D. 1799* (U•X•L, 2002), as well as extensive contributions to Gale's seven-volume *Science and Its Times* (2000-2001). As a writer on history, Knight has published *Middle Ages Reference Library* (2000), *Ancient Civilizations* (1999), and a volume in U•X•L's *African American Biography* series (1998). Knight's publications in the realm of music include *Parents Aren't Supposed to Like It* (2001), an overview of contemporary performers and genres, as well as *Abbey Road to Zapple Records: A Beatles Encyclopedia* (Taylor, 1999).

COMMENTS AND SUGGESTIONS

Your comments on this series and suggestions for future editions are welcome. Please write: The Editor, *Science of Everyday Things,* Gale Group, 27500 Drake Road, Farmington Hills, MI 48331-3535.

ADVISORY BOARD

UNDERSTANDING THE EARTH SCIENCES

EARTH, SCIENCE, AND NONSCIENCE

GEOSCIENCE AND EVERYDAY LIFE

EARTH SYSTEMS

EARTH, SCIENCE, AND NONSCIENCE

CONCEPT

To understand the composition and structure of Earth, one must comprehend the forces that shaped it. Much the same is true of the earth sciences themselves, which originated from attempts to explain the origins of Earth and the materials of which it is composed. Before the modern era, such explanations had roots in religion, mythology, or philosophy and drew from preconceived ideas rather than from observed data. A turning point came with the development of the scientific method, a habit of thinking that spread from astronomy and physics to chemistry and the earth sciences.

HOW IT WORKS

ARISTOTLE'S FOUR CAUSES

Though the Greek philosopher Aristotle (384–322 B.C.) exerted a negative influence on numerous aspects of what became known as the physical sciences (astronomy, physics, chemistry, and the earth sciences), he is still rightly regarded as one of the greatest thinkers of the Western world. Among his contributions to thought was the identification of four causes, or four approaches to the question of how and why something exists as it does.

In Aristotle's system, which developed from ideas of causation put forward by his predecessors, the most basic of explanations is the *material cause,* or the substance of which a thing is made. In a house, for instance, the wood and other building materials would be the material cause. The builders themselves are the *efficient cause,* or the forces that shaped the house. More complex than these is a third variety of cause-effect relationship, the *formal cause*—that is, the design or blueprint on which something is modeled.

The first three Aristotelian causes provide a pathway for explaining *how;* the fourth and last cause approaches the much more challenging question of *why.* This is the *final cause,* or the reason why a thing exists at all—in other words, the purpose for which it was made. Even in the case of the house, this is a somewhat complicated matter. A house exists, of course, to provide a dwelling for its occupants, but general contractors would not initiate the building process if they did not expect to make a profit, nor would the subcontractors and laborers continue to work on it if they did not earn an income from the project.

RELIGION, SCIENCE, AND EARTH

The matter of final cause is almost unimaginably more complex when applied to Earth rather than to a house. The question "Why does Earth exist?" or "What is the ultimate reason for Earth's existence?" is not really a topic for science at all, but rather for theology and philosophy. Nor do the answers provided by religion and philosophical beliefs qualify as answers in the same sense that workable scientific theories do.

There has always been a degree of tension between religion and the sciences, and nowhere has this been more apparent than in the earth sciences. As will be discussed later in this essay, most early theories concerning Earth's structure and development were religious in origin, and even some modern explanations have theological roots. Certainly there is nothing wrong with a

ENGRAVING AFTER A MARBLE BUST OF ARISTOTLE. *(Library of Congress.)*

scientist having religious beliefs, as long as those beliefs do not provide a filter for all data. If they do, the theologically minded scientist becomes rather like a mathematician attempting to solve a problem on the basis of love rather than reason. Most people would agree that love is higher and greater than mathematics; nonetheless, it has absolutely no bearing on the subject.

SCIENTIFIC ANSWERS AND THE SEARCH FOR A DESIGNER. The third, or formal, cause is less fraught with problems than the final cause when applied to the study of Earth, yet it also illustrates the challenges inherent in keeping science and theology separate. Does Earth have a "design," or blueprint? The answer is yes, no, and maybe. Yes, Earth has a design in the sense that there is an order and a balance between its components, a subject discussed elsewhere with reference to the different spheres (geosphere, hydrosphere, biosphere, and atmosphere). The physical evidence, however, tends to suggest a concept of design quite different from the theistic notion of a deity who acts as creator.

Consider, for example, the ability of an animal to alter its appearance as a means of blending in with its environment, to ward off predators, to disguise itself while preying upon other animals, or for some other purpose. On the one hand, this seems like an example of conscious design by a loving creator, but as Charles Darwin (1809–1882) showed, it may simply be a matter of adaptability. According to Darwin, members of species unable to alter their appearance died out, leading to the dominance of those who could camouflage themselves.

In fact, science is not really capable of addressing the matter of a Designer (i.e., God), and thus, for scientists, the question of a deity's role in nature is simply irrelevant. This is not because scientists are necessarily atheists (many are and have been dedicated men and women of faith) but because the concept of a deity simply adds an unnecessary step to scientific analysis.

This is in line with Ockham's razor, a principle introduced by the medieval English philosopher William of Ockham (*ca.* 1285?–1349). According to Ockham, "entities must not be unnecessarily multiplied." In other words, in analyzing any phenomenon, one should seek the simplest and most straightforward explanation. Scientists are concerned with hard data, such as the evidence obtained from rock strata. The application of theological ideas in such situations would at best confuse and complicate the process of scientific analysis.

THE ARGUMENT FROM DESIGN. A few years before Ockham, the Italian philosopher Thomas Aquinas (1224 or 1225–1274) introduced a philosophical position known as the "argument from design." According to Aquinas, whose idea has been embraced by many up to the present day, the order and symmetry in nature indicate the existence of God. Some philosophers have conceded that this order does indeed indicate the existence of *a* god, though not necessarily the God of Christianity. Science, however, cannot afford to go even that far: where spiritual matters are concerned, science must be neutral.

Does any of this disprove the existence of God? Absolutely not. Note that science must be *neutral,* not in opposition, where spiritual matters are concerned. Indeed, one could not disprove God's existence scientifically if one wanted to do so; to return to the analogy given earlier, such an endeavor would be akin to using mathematics to disprove the existence of love. Religious matters are simply beyond the scope of science, and to use science against religion is as misinformed a position as its opposite.

SCIENCE AND THE FIRST TWO CAUSES. To return to Aristotle's causes, let us briefly consider the material and efficient cause as applied to the subject of Earth. These are much simpler matters than formal and final cause, and science is clearly able to address them. An understanding of Earth's material cause—that is, its physical substance—requires a brief examination of the chemical elements. The elements are primarily a subject for chemistry, though they are discussed at places throughout this book, inasmuch as they relate to the earth sciences and, particularly, geochemistry. Furthermore, the overall physical makeup of Earth, along with particular aspects of it, are subjects treated in much greater depth within numerous essays concerning specific topics, such as sedimentation or the biosphere.

Likewise the efficient cause, or the complex of forces that have shaped and continue to shape Earth, is treated in various places throughout this book. In particular, the specifics of Earth's origins and the study of these origins through the earth sciences are discussed in essays on aspects of historical geology, such as stratigraphy. Here the origins of Earth are considered primarily from the standpoint of the historical shift from mythological or religious explanations to scientific ones.

REAL-LIFE APPLICATIONS

MYTHOLOGY AND GEOLOGY

Most of what people believed about the origins and makeup of Earth before about 1700 bore the imprint of mythology or merely bad science. Predominant among these theories were the Creation account from the biblical Book of Genesis and the notion of the four elements inherited from the Greeks. These four elements—earth, air, fire, and water—were said to form the basis for the entire universe, and thus every object was thought to be composed of one or more of these elements. Thanks in large part to Aristotle, this belief permeated (and stunted) the physical sciences.

To call the biblical Creation story mythology is not, in this context at least, a value judgment. The Genesis account is not scientific, however, in the sense that it was not written on the basis of observed data but rather from religious principles. The concept of the four elements at least relates somewhat to observation, but specifically to untested observation; for this reason, it is hardly more scientific than the Genesis Creation story. The four elements were not, strictly speaking, a product of mythology, but they were mythological in the pejorative sense—that is, they had no real basis in fact.

GEOMYTHOLOGY. The biblical explanation of Earth's origins is but one of many creation myths, part of a larger oral and literary tradition that Dorothy B. Vitaliano, in her 1973 book *Legends of the Earth,* dubbed *geomythology.* Examples of geomythology are everywhere, and virtually every striking natural feature on Earth has its own geomythological backdrop. For instance, the rocky outcroppings that guard the western mouth of the Mediterranean, at Gibraltar in southern Spain and Ceuta in northern Morocco, are known collectively as the Pillars of Hercules because the legendary Greek hero is said to have built them.

Geomythological stories can be found in virtually all cultures. For instance, traditional Hawaiian culture explains the Halemaumau volcano, which erupted almost continuously from 1823 to 1924, as the result of anger on the part of the Tahitian goddess Pele. Native Americans in what is now Wyoming passed down legends concerning the grooves along the sides of Devils Tower, which they said had been made by bears trying to climb the sides to escape braves hunting them.

GREEK GEOMYTHOLOGY. In Western culture, among the most familiar examples of geomythology, apart from those in the Bible, are the ones that originated in ancient Greece and Rome. The Pillars of Hercules represents but one example. In particular, the culture of the Greeks was infused with geomythological elements. They believed, for instance, that the gods lived on Mount Olympus and spoke through the Delphic Oracle, a priestess who maintained a trancelike state by inhaling intoxicating vapors that rose through a fault in the earth.

Much of Greek mythology is actually geomythology. Most of the principal Greek deities ruled over specific aspects of the natural world that are today the province of the sciences, and many of them controlled realms now studied by the earth sciences and related disciplines. Certain branches of geology today are concerned with Earth's interior, which the Greeks believed was controlled by Hades, or the Roman god Pluto. Volcanoes and thunderbolts were the work of the blacksmith god Hephaestus (the Roman deity Vulcan), while Poseidon (known to the Romans as Neptune) oversaw the area studied today by oceanographers.

ATLANTIS. Among the most persistent geomyths with roots in Greek civilization is the story of Atlantis, a continent that allegedly sank into the sea. Over the years, the myth grew to greater and greater dimensions, and in a blurring between the Atlantis myth and the biblical story of Eden, Atlantis came to be seen as a lost utopia. Even today, some people believe in Atlantis, and for scholarly endorsement they cite a passage in the writings of Plato (427?–347 B.C.). The great Greek philosopher depicted Atlantis as somewhere beyond the Pillars of Hercules, and for this reason its putative location eventually shifted to the middle of the Atlantic—an ocean in fact named for the "lost continent."

Given the layers of mythology associated with Atlantis, it may come as a surprise that the story has a basis in fact and that accounts of it appear in the folklore of peoples from Egypt to Ireland. It is likely that the myth is based on a cataclysmic event, either a volcanic eruption or an earthquake, that took place on the island of Crete, as well as nearby Thíra, around 1500 B.C. This cataclysm, some eight centuries before the rise of classical Greek civilization, brought an end to the Minoan civilization centered around Knossos in Crete. Most likely it raised vast tidal waves, or tsunamis, that reached lands far away and may have caused other cities or settlements to disappear beneath the sea.

BIBLICAL GEOMYTHOLOGY. As important as such Greek stories are, no geomythological account has had anything like the impact on Western civilization exerted by the first nine chapters of the Bible. These chapters contain much more than geomythology, of course; in fact, they introduce the central themes of the Bible itself: righteousness, sin, redemption, and God's covenant with humankind. In these nine chapters (or, more properly, eight and a half chapters), which cover the period from Earth's creation until the Great Flood, events are depicted as an illustration of this covenant. Thus, in 9 Genesis, when God introduces the rainbow after the Flood, he does so with the statement that it is a sign of his promise never again to attempt to destroy humanity.

As with Atlantis, the story of the Great Flood appears in other sources as well. Its antecedents include the Sumerian Gilgamesh epic, which originated in about 2000 B.C., a millennium before the writing of the biblical account. Also as in the case of Atlantis, the biblical flood seems to have a basis in fact. Some modern scientists theorize that the Black Sea was once a freshwater lake, until floods covered the land barriers that separated it from saltwater.

The Flood occupies chapters 6 through 9 of Genesis, while chapters 3 through 5 are concerned primarily with human rather than geologic events. The story of Adam, Eve, the serpent, and the fruit of the Tree of Knowledge is a beautiful, complex, and richly symbolic explanation of how humans, born innocent, are prone to sin. It is the first conflict between God and human, just as Cain's murder of Abel is the first conflict between people. Both stories serve to illustrate the themes mentioned earlier: in both cases, God punishes the sins of the humans but also provides them with protection as a sign of his continued faithfulness.

THE BIBLE AND SCIENCE. In fact, the entire Creation story, source of centuries' worth of controversy, occupies only two chapters, and this illustrates just how little attention the writers of the Bible actually devoted to "scientific" subjects. Certainly, many passages in the Bible describe phenomena that conflict with accepted scientific knowledge, but most of these fall under the classification of miracles—or, if one does not believe them, alleged miracles. Was Jesus born of a virgin? Did he raise the dead? People's answers to those questions usually have much more to do with their religious beliefs than with their scientific knowledge.

Most of the biblical events related to the earth sciences appear early in the Old Testament, and most likewise fall under the heading of "miracles." Certain events, such as the parting of the Red Sea by Moses, even have possible scientific explanations: some historians believe that there was actually an area of dry land in the Red Sea region and that Moses led the children of Israel across it. The account of Joshua causing the Sun to stand still while his men marched around the city of Jericho is a bit more difficult to square with science, but a believer might say that the Sun (or rather, Earth) *seemed* to stand still.

In any case, the Bible does not present itself as a book of science, and certainly the Israelites of ancient times had little concept of science as we know it today. Some of the biblical passages mentioned here have elicited controversy, but few have inspired a great deal of discussion, precisely because they are generally regarded as accounts of miracles. The same is not true, however, of the first two chapters of Genesis, which even today remain a subject of dispute in some quarters.

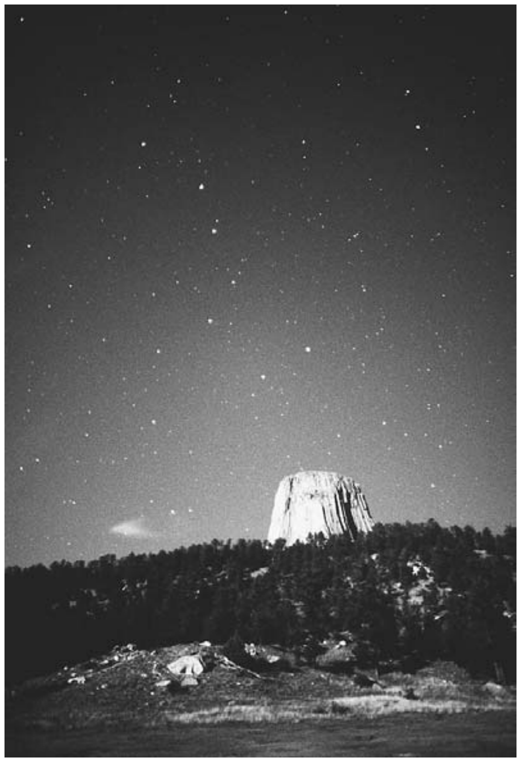
SIX DAYS? Actually, 2 Genesis concerns Adam's life before the Fall as well as the creation of Eve from his rib, so the Creation story proper is confined to the first chapter. One of the most famous passages in Western literature, 1 Genesis describes God's creation of the universe in all its particulars, each of which he spoke into being, first by saying, "Let there be light." After six days of activity that culminated with the creation of the human being, he rested, thus setting an example for the idea of a Sabbath rest day.

As prose poetry, the biblical Creation story is among the great writings of all time. It is also a beautiful metaphoric description of creation by a loving deity; but it is not a guide to scientific study. Yet for many centuries, Western adherence to the Genesis account (combined with a number of other factors, including the general stagnation of European intellectual life throughout much of the medieval period) forced a virtual standstill of geologic study. The idea that Earth was created in 144 hours reached its extreme with the Irish bishop James Ussher (1581–1656), who, using the biblical genealogies from Adam to Christ, calculated that God finished making Earth at 9:00 A.M. on Sunday, October 23, 4004 B.C.

THE MYTH OF THE FOUR ELEMENTS

Religion alone is far from the only force that has slowed the progress of science over the years. Sometimes the ideas of scientists or philosophers themselves, when formed on the basis of something other than scientific investigation, can prove at least as detrimental to learning. Such is the case when thinkers become more dedicated to the theory than to the pursuit of facts, as many did in their adherence to the erroneous concept of the four elements.

Today scientists understand an element as a substance made up of only one type of atom, meaning that unlike a compound, it cannot be broken down chemically into a simpler substance. This definition developed over the period from about 1650 to 1800, thanks to the British chemist Robert Boyle (1627–1691), who originated the idea of elements as the simplest substances; the French chemist Antoine Lavoisier (1743–1794), who first distinguished between elements and compounds; and the British chemist John Dalton (1766–1844), who introduced the atomic theory of matter.



DEVILS TOWER, WITH THE BIG DIPPER VISIBLE IN THE NIGHT SKY. *(© Jerry Schad/Photo Researchers. Reproduced by permission.)*

During the twentieth century, with the discovery of the atomic nucleus and the protons within it, scientists further refined their definition of an element. Today elements are distinguished by atomic number, or the number of protons in the atomic nucleus. Carbon, for instance, has an atomic number 6, meaning that there are six protons in the carbon nucleus; therefore, any element with six protons in its atomic nucleus *must* be carbon.

ATOMIC THEORY VERSUS THE FOUR ELEMENTS. Atomic, or corpuscular, theory had been on the rise for some 150 years before Dalton, who built on ideas of predecessors that included Galileo Galilei (1564–1642) and Sir Isaac Newton (1642–1727). In any case, the first thinker to conceive of atoms lived more than 2,000 years earlier. He was Democritus (*ca.* 460–*ca.* 370 B.C.), a Greek philosopher who described the world as being composed of indivisible particles—*atomos* in Greek. Democritus's idea was far from modern scientific atomic theory, but it came much closer than any other theory before the Scientific Revolution (*ca.* 1550–1700).

Why, then, did it take so long for Western science to come around to the atomic idea? The answer is that Aristotle, who exerted an almost incalculable impact on Muslim and Western thought during the Middle Ages, rejected Democritus' atomic theory in favor of the four elements theory. The latter had its roots in the very beginnings of Greek ideas concerning matter, but it was the philosopher Empedocles (*ca.* 490–430 B.C.) who brought the notion to some kind of maturity.

A NONSCIENTIFIC THEORY. According to the four elements theory, every object could be identified as a combination of elements: bone, for instance, was supposedly two parts earth, two parts water, and two parts fire. Of course, this is nonsense, and, in fact, none of the four elements are even really elements. Water comes the closest, being a compound of the elements hydrogen and oxygen. Earth and air are mixtures, while fire is the result of combustion, a form of oxidation-reduction chemical reaction.

Nonetheless, the theory had at least some basis in observation, since much of the physical world seems to include liquids, things that grow from the ground, and so on. Such observations alone, of course, are not enough to construct a theory, as would have become apparent if the Greeks had attempted to test their ideas. The ancients, however, tended to hold scientific experimentation in low esteem, and they were more interested in applying their intellects to the development of ideas than they were in getting their hands dirty by putting their concepts to the test.

THINKING IN FOURS. Aristotle explained the four elements as combinations of four qualities, or two pairs of opposites: hot/cold and wet/dry. Thus, fire was hot and dry, air was dry and cold, water was cold and wet, and earth was wet and hot. It is perhaps not accidental that there were four elements, four qualities, or even perhaps four Aristotelian causes.

Much earlier, the philosopher and mathematician Pythagoras (*ca.* 580–*ca.* 500 B.C.), who held that all of nature could be understood from the perspective of numbers, first suggested the idea of four basic elements because, he maintained, the number four represents perfection. This concept influenced Greek thinkers, including Empedocles and even Aristotle, and is also probably the reason for the expression *four corners of the world.*

That expression, which conveys a belief in a flat Earth, raises an important point that must be made in passing. Despite his many erroneous ideas, Aristotle was the first to prove that Earth is a sphere, which he showed by observing the circular shadow on the Moon during a lunar eclipse. This points up the fact that ancient thinkers may have been misguided in many regards, yet they still managed to make contributions of enormous value. In the same vein, Pythagoras, for all his strange and mystical ideas, greatly advanced scientific knowledge by introducing the concept that numbers can be applied to the study of nature.

In any case, the emphasis on fours trickled down through classical thought. Thus, the great doctors Hippocrates (*ca.* 460–*ca.* 377 B.C.) and Galen (129–*ca.* 199) maintained that the human body contains four "humors" (blood, black bile, green bile, and phlegm), which, when imbalanced, caused diseases. Humoral theory would exert an incalculable toll on human life throughout the Middle Ages, resulting in such barbaric medical practices as the use of leeches to remove "excess" blood from a patient's body. The idea of the four elements had a less clearly pernicious effect on human well-being, yet it held back progress in the sciences and greatly impeded thinkers' understanding of astronomy, physics, chemistry, and geology.

The Showdown Between Myth and Science

Aristotle's teacher Plato had accepted the idea of the four elements, but proposed that space is made up of a fifth, unknown element. This meant that Earth and the rest of the universe are fundamentally different, a misconception that prevailed for two millennia. Aristotle adopted that idea, as well as Plato's concept of a Demiurge, or Prime Mover, as Aristotle called it. Centuries later Aquinas equated Aristotle's Prime Mover with the Christian God.

Building on these and other ideas, Aristotle proceeded to develop a model of the cosmos in which there were two principal regions: a celestial, or heavenly, realm above the orbit of the Moon and a terrestrial, or earthly, one in what was known as the sublunary (below the Moon) region. Virtually everything about these two realms differed. The celestial region never changed, whereas change was possible on Earth. Earth itself consisted of the four elements, whereas the heavens were made up of a fifth substance, which he called *ether.*

If left undisturbed, Aristotle theorized, the four elements would completely segregate into four concentric layers, with earth at the center, surrounded by water, then air, and then fire, bounded at the outer perimeter by the ether. The motion of bodies above the Moon's sphere caused the elements to behave unnaturally, however, and thus they remained mixed and in a constant state of agitation.

The distinction between so-called natural and unnatural (or violent) motion became one of the central ideas in Aristotle's physics, a scientific discipline whose name he coined in a work by the same title. According to Aristotle, all elements seek their natural position. Thus, the element earth tends to fall toward the center of the universe, which was identical with the center of Earth itself.

THE SCIENTIFIC REVOLUTION. On these and other ideas, Aristotle built a complex, systematic, and almost entirely incorrect set of principles that dominated astronomy and physics as well as what later became the earth sciences and chemistry. The influence of Aristotelian ideas on astronomy, particularly through the work of the Alexandrian astronomer Ptolemy (*ca.* 100–170), was especially pronounced.

It was through astronomy, the oldest of the physical sciences, that the Aristotelian and Ptolemaic model of the physical world ultimately was overthrown. This revolution began with the proof, put forward by Nicolaus Copernicus (1473–1543), that Earth is not the center of the universe. The Catholic Church, which had controlled much of public life in Europe for the past thousand years, had long since accepted Ptolemy's geocentric model on the reasoning that if the human being is created in God's image, Earth must be at the center of the universe. Copernicus' heliocentric (Sun-centered) cosmology therefore constituted a challenge to religious authority—a very serious matter at a time when the Church held the power of life and death.

Copernicus died before he suffered the consequences of his ideas, but Galileo, who lived much later, found himself in the middle of a debate between the Church and science. This conflict usually is portrayed in simplistic terms, with Galileo as the noble scientific genius defending reason against the powers of reaction, but the facts are much more complex. For centuries, the Church had preserved and encouraged learning, and the reactionary response to Copernican ideas must be understood in light of the challenges to Catholic authority posed by the Protestant Reformation. Furthermore, Galileo was far from diplomatic in his dealings, for instance, deliberately provoking Pope Urban VIII (1568–1644), who had long been a friend and supporter.

In any case, Galileo made a number of discoveries that corroborated Copernicus' findings while pointing up flaws in the ideas of Aristotle and Ptolemy. He also conducted studies on falling objects that, along with the laws of planetary motion formulated by Johannes Kepler (1571–1630), provided the basis for Newton's epochal work in gravitation and the laws of motion. Perhaps most of all, however, Galileo introduced the use of the scientific method.

THE SCIENTIFIC METHOD. The scientific method is a set of principles and procedures for systematic study using evidence that can be clearly observed and tested. It consists of several steps, beginning with observation. This creates results that lead to the formation of a hypothesis, an unproven statement about the way things are. Up to this point, we have gone no further than ancient science: Aristotle, after all, was making a hypothesis when he said, for instance, that heavy objects fall faster than light ones, as indeed they seem to do.

Galileo, however, went beyond the obvious, conducting experiments that paved the way for modern understanding of the acceleration due to gravity. As it turns out, heavy objects fall faster than light ones only in the presence of resistance from air or another medium, but in a vacuum a stone and a feather would fall at the same rate. How Galileo arrived at this idea is not important

KEY TERMS

ATOM: The smallest particle of an element, consisting of protons, neutrons, and electrons. An atom can exist either alone or in combination with other atoms in a molecule.

ATOMIC NUMBER: The number of protons in the nucleus of an atom.

COMPOUND: A substance made up of atoms of more than one element, chemically bonded to one another.

COSMOLOGY: A branch of astronomy concerned with the origin, structure, and evolution of the universe.

COSMOS: The universe.

ELEMENT: A substance made up of only one kind of atom. Unlike compounds, elements cannot be broken chemically into other substances.

GEOCENTRIC: Earth-centered.

GEOMYTHOLOGY: Folklore inspired by geologic phenomena.

HELIOCENTRIC: Sun-centered.

HYPOTHESIS: An unproven statement regarding an observed phenomenon.

LAW: A scientific principle that is shown always to be the case and for which no exceptions are deemed possible.

PHYSICAL SCIENCES: Astronomy, physics, chemistry, and the earth sciences.

PROTON: A positively charged particle in an atom.

SCIENTIFIC METHOD: A set of principles and procedures for systematic study that includes observation; the formation of hypotheses, theories, and ultimately laws on the basis of such observation; and continual testing and reexamination.

SCIENTIFIC REVOLUTION: A period of accelerated scientific discovery that completely reshaped the world. Usually dated from about 1550 to 1700, the Scientific Revolution saw the origination of the scientific method and the introduction of such ideas as the heliocentric (Sun-centered) universe and gravity.

THEORY: A general statement derived from a hypothesis that has withstood sufficient testing.

VACUUM: An area devoid of matter, even air.

here; rather, his application of the scientific method, which requires testing of hypotheses, is the key point.

If a hypothesis passes enough tests, it becomes a theory, or a general statement. An example of a theory is uniformitarianism, an early scientific explanation of Earth's origins discussed elsewhere, in the context of historical geology. Many scientific ideas remain theories and are quite workable as such: in fact, much of modern physics is based on the quantum model of subatomic behavior, which remains a theory. But if something always has been observed to be the case and if, based on what scientists know, no exceptions appear possible, it becomes a law. An example is Newton's third law of motion: no one has ever observed or created a situation in which a physical action does not yield an equal and opposite reaction.

Even laws can be overturned, however, and every scientific principle therefore is subjected to continual testing and reexamination, making the application of the scientific method a cyclical process. Thus, to be scientific, a principle must be capable of being tested. It should also be said that one of the hallmarks of a truly scientific theory is the attitude of its adherents. True scientists are always attempting to *disprove* their own ideas by subjecting them to rigorous tests; the more such tests a theory survives, the stronger it becomes.

CREATIONISM: RELIGION UNDER A VEIL OF SCIENCE

During the twentieth century, a movement called creationism emerged at the fringes of science. Primarily American in origin, creationism is a fundamentalist Christian doctrine, meaning that it is rooted in a strict literal interpretation of the Genesis account of Creation. (For this reason, creationism has little influence among Christians and Christian denominations not prone to literalism.) From the 1960s onward, it has been called *creation science,* but even though creationism sometimes makes use of scientific facts, it is profoundly unscientific.

Again, the reference to creationism as unscientific does not necessarily carry a pejorative connotation. Many valuable things are unscientific; however, to call creationism unscientific is pejorative in the sense that its adherents *claim* that it is scientific. The key difference lies in the attitude of creationists toward their theory that God created the Earth if not in six literal days, then at least in a very short time.

If this were a genuine scientific theory, its adherents would be testing it constantly against evidence, and if the evidence contradicted the theory, they would reject the theory, not the evidence. Science begins with facts that lead to the development of theories, but the facts always remain paramount. The opposite is true of creationism and other nonscientific beliefs whose proponents simply look for facts to confirm what they have decided is truth. Conflicting evidence simply is dismissed or incorporated into the theory; thus, for instance, fossils are said to be the remains of animals who did not make it onto Noah's ark.

Creationism (for which *The Oxford Companion to the Earth* provides a cogent and balanced explanation) is far from the only unscientific theory that has pervaded the hard sciences, the social sciences, or society in general. Others, aside from the four elements, have included spontaneous generation and the phlogiston theory of fire as well as various bizarre modern notions, such as flat-Earth theory, Holocaust or Moon-landing denial, and Afrocentric views of civilization as a vast racial conspiracy. Compared with Holocaust denial, for instance, creationism is benign in the sense that its proponents seem to act in good faith, believing that any challenge to biblical literalism is a challenge to Christianity itself.

Still, there is no justification for the belief that Earth is very young; quite literally, mountains of evidence contradict this claim. Nor is the idea of an old Earth a recent development; rather, it has circulated for several hundred years—certainly long before Darwin's theory of evolution, the scientific idea with which creationists take the most exception. For more about early scientific ideas concerning Earth's age, see Historical Geology and essays on related subjects, including Paleontology and Geologic Time. These essays, of course, are concerned primarily with modern theories regarding Earth's history, as well as the observations and techniques that have formed the basis for such theories. They also examine pivotal early ideas, such as the Scottish geologist James Hutton's (1726–1797) principle of uniformitarianism.

WHERE TO LEARN MORE

Bender, Lionel. *Our Planet.* New York: Simon and Schuster Books for Young Readers, 1992.

Elsom, Derek M. *Planet Earth.* Detroit: Macmillan Reference USA, 2000.

Gamlin, Linda. *Life on Earth.* New York: Gloucester Press, 1988.

Hancock, Paul L., and Brian J. Skinner. *The Oxford Companion to the Earth*. New York: Oxford University Press, 2000.

Llamas Ruiz, Andrés. *The Origin of the Universe.* Illus. Luis Rizo. New York: Sterling Publishers, 1997.

Skinner, Brian J., Stephen C. Porter, and Daniel B. Botkin. *The Blue Planet: An Introduction to Earth System Science.* 2nd ed. New York: John Wiley and Sons, 1999.

The Talk. Origins Archive: Exploring the Creation/Evolution Controversy (Web site). <http://www.talkorigins.org/>.

Van der Pluijm, Ben A., and Stephen Marshak. *Earth Structure* (Web site). <http://www-personal.umich.edu/~vdpluijm/earthstructure.htm>.

Vitaliano, Dorothy B. *Legends of the Earth*. Bloomington: Indiana University Press, 1973.

Web Elements (Web site). <http://www.webelements.com/>.

Windows to the Universe (Web site). <http://www.windows.ucar.edu/win_entry.html>.

GEOSCIENCE AND EVERYDAY LIFE

CONCEPT

How can learning about rocks help us in our daily lives? The short answer is that geology and the related geologic sciences (sometimes referred to collectively as *geoscience*) give us a glimpse of the great complexity inherent in the natural world, helping us appreciate the beauty and order of things. This, in turn, makes us aware of our place in the scheme of things, so that we begin to see our own daily lives in their proper context. Beyond that, the study of geoscientific data can give us an enormous amount of information of practical value while revealing much about the world in which we dwell. The earth sciences are, quite literally, all around us, and by learning about the structures and processes of our planet, we may be surprised to discover just how prominent a place geoscience occupies in our daily lives and even our thought patterns.

HOW IT WORKS

Why Study Geoscience?

One of the questions students almost always ask themselves or their teachers is "How will I use this?" or "What does all this have to do with everyday life?" It is easy enough to understand the application of classes involved in learning a trade or practical skill—for example, wood shop or a personal finance course. But the question of applicability sometimes becomes more challenging when it comes to many mathematical and scientific disciplines. Such is the case, for instance, with the earth sciences and particularly geoscience. Yet if we think about these concerns for just a moment, it should become readily apparent just why they are applicable to our daily lives.

After all, geoscience *is* the study of Earth, and therefore it relates to something of obvious and immediate practical value. We may think of a hundred things more important and pressing than studying Earth—romantic involvements, perhaps, or sports, or entertainment, or work (both inside and outside school)—yet without Earth, we would not even have those concerns. Without the solid ground beneath our feet, which provides a stage or platform on which these and other activities take place, life as we know it would be simply impossible. Our lives are bounded by the solid materials of Earth—rocks, minerals, and soil—while our language reflects the primacy of Earth in our consciousness. As we discuss later, everyday language is filled with geologic metaphors.

Defining Geoscience

The geologic sciences—geology, geophysics, geochemistry, and related disciplines—are sometimes referred to together as geoscience. They are united in their focus on the solid earth and the mostly nonorganic components that compose it. In this realm of earth science, geology is the leading discipline, and it has given birth to many offshoots, including geophysics and geochemistry, which represent the union of geology with physics and chemistry, respectively.

Geology is the study of the solid earth, especially its rocks, minerals, fossils, and land formations. It is divided into historical geology, which is concerned with the processes whereby Earth was formed, and physical geology, or the study of the materials that make up the planet. Geophysics addresses Earth's physical processes as well as its gravitational, magnetic, and electric properties and the means by which energy is transmitted through its interior. Geochemistry is concerned with the chemical properties and processes of Earth—in particular, the abundance and interaction of chemical elements.

These subjects are of principal importance in this book. Though geology takes the lion's share of attention, geophysics and geochemistry each encompass areas of study essential to understanding our life on Earth: hence we look in separate essays at such geophysical subjects as Gravity and Geodesy or Geomagnetism as well as such geochemical topics as Biogeochemical Cycles, Carbon Cycle, and Nitrogen Cycle.

OTHER AREAS OF GEOSCIENCE. In addition to these principal areas of interest in geoscience, this book treats certain subdisciplines of geology as areas of interest in their own right. These include geomorphology and the studies of sediment and soil. Geomorphology is an area of physical geology concerned with the study of landforms, with the forces and processes that have shaped them, and with the description and classification of various physical features on Earth.

In contrast to geology, which normally is associated with rocks and minerals, geomorphology is concerned more with larger configurations, such as mountains, or with the erosive and weathering forces that shape such landforms. (See, for instance, essays on Mountains, Erosion, and Mass Wasting.) Erosion and weathering also play a major role in creating sediment and soil, areas that are of interest in the subdisciplines of sedimentology and soil science.

CONTRAST WITH OTHER DISCIPLINES AND SUBDISCIPLINES. Geoscience is distinguished sharply from the other branches of the earth sciences, namely, hydrologic sciences and atmospheric sciences. The first of these sciences, which is concerned with water, receives attention in essays on Hydrology and Hydrologic Cycle. The second, which includes meteorology (weather forecasting) and climatology, is the subject of the essays Weather and Climate.

In addition to the hydrologic and atmospheric sciences, there are areas of earth sciences study that touch on biology. Essays in this book that treat biosphere-related topics include Ecosystems and Ecology and Ecological Stress. There is one area or set of areas, however, in which geoscience and biology more or less overlap: sedimentology and soil science, since soil is a combination of rock fragments and organic material (see Soil).

The Territory of Geoscience

The organic material in soil—dead plants and animals and parts thereof—has ceased to be part of the biosphere and is part of the geosphere. The geosphere encompasses the upper part of Earth's continental crust, or that portion of the solid earth on which human beings live and which provides them with most of their food and natural resources. (For more about the "spheres," see Earth Systems.)

Later in this essay, we discuss several areas of geoscientific study that take place close to the surface of Earth. Yet the territory of geoscience extends far deeper, going well below the geosphere into the interior of the planet. (For more on this subject, see Earth's Interior.) Geoscience even involves the study of "earths" other than our own; as discussed in such essays as Planetary Science and Sun, Moon, and Earth, there is considerable overlap between geoscience and astronomy.

REAL-LIFE APPLICATIONS

The Primacy of Earth

We may not think about geoscience or earth science much, or at least we may not *think* that we think about these topics very much—and yet we spend our lives in direct contact with these areas. Certainly in a given day, every person experiences physics (the act of getting out of bed is an example of the third law of motion, discussed in Gravity and Geodesy) and chemistry (eating and digesting food, for instance), but the experience of geoscience is more direct: we can actually *touch* the earth.

Before the late nineteenth century and the introduction of processed foods, everything a person ate clearly either was grown in the soil or was part of an animal that had fed on plants grown in the soil. Even today, the most grotesquely processed products, such as the synthetic cream puffs sold at a convenience store, still hold a connection to the earth, inasmuch as they contain sugar—a natural product. In any case, most of what we eat (especially in a health-conscious diet) has a close connection to the earth.

GEOSCIENCE AND LANGUAGE. No wonder, then, that a number of creation stories, including the one in Genesis, depict humankind as coming from the soil—an account of origins reflected in the well-known graveside benediction "Ashes to ashes, dust to dust." Our language is filled with geoscientific metaphors, including such proverbs as "A rolling stone gathers no moss" or "Still waters run deep." (The latter aphorism, despite its hydrologic imagery, actually refers to the fact that in deeper waters, rock formations are, by definition, not likely to be near the surface. By contrast, in order for a "babbling brook" to make as much noise as it does, it must be flowing over prominent rocks.)

Then there are the countless geologic figures of speech: "rock solid," "making mountains out of molehills," "cold as a stone," and so on. When the rock musician Bob Seger sang, in a 1987 hit, about being "Like a Rock" as a younger man, listeners knew exactly what he meant: solid, strong, dependable. So established was the metaphor that a few years later, Chevrolet used the song in advertising their trucks and sport-utility vehicles (including, ironically, a vehicle whose name uses a somewhat less reassuring geologic image: the Chevy Avalanche).

THE GEOMORPHOLOGY OF RELIGIOUS FAITH. Rocks and other geologic features have long captured the imagination of humans; hence, we have the many uses of mountains in, for instance, religious imagery. There was the mystic mountain paradise of Valhalla in Norse mythology as well as Mount Olympus in Greek myths and legends. Unlike Valhalla, Olympus is a real place; so, too, is Kailas in southwestern Tibet, which ancient adherents of the Jain religion called Mount Meru, the center of the cosmos, and which Sanskrit literature identifies as the paradise of Siva, one of the principal Hindu deities.

There is also Sri Pada, or Adam's Peak, in Sri Lanka, a spot sacred to four religions. Buddhists believe the mountain is the footprint of the Buddha, while Hindus call it the footprint of Siva. Muslims and Christians believe it to be the footprint of Adam. Then there are the countless mountain locales of the Old Testament, including Ararat (in modern Turkey), where Noah's ark ran aground, and Sinai (in the Sinai Desert between Egypt and Israel), where Moses was called by God and later received the Ten Commandments.

The New Testament account of the life of Jesus Christ is punctuated throughout with geologic and geomorphologic details: the temptations in the desert, the Sermon on the Mount, and the Transfiguration, which probably took place atop Mount Tabor in Israel. He was crucified on a hill, buried in a cave, rolled a stone away at his Resurrection, and finally ascended to heaven from the Mount of Olives.

Arts, Media, and the Geosciences

From ancient times rocks and minerals have intrigued humans, not only by virtue of their usefulness but also because of their beauty. On one level there is the purely functional use of rock as a building material, and on another level there is the aesthetic appreciation for the beauty imparted by certain types of rock, such as marble.

Rock is an excellent building material when it comes to compression, as exerted by a great weight atop the rock; in the case of tension or stretching, however, rock is very weak. This shortcoming of stone, which was otherwise an ideal building material for the ancients (given its cheapness and relative abundance in some areas of the world), led to one of history's great innovations in architecture and engineering: the arch. A design feature as important for its aesthetic value as for its strength, the arch owed its physical power to the principle of weight redistribution. Arched Roman structures two thousand or more years old still stand in Europe, a tribute to the interaction of art, functionality, and geoscience.

THE VISUAL ARTS. *The Oxford Companion to the Earth* contains a number of excellent entries on the relationship between geoscience and the arts. In the essay "Art and the Earth Sciences," for instance, Andrew C. Scott notes four ways in which the earth sciences and the visual arts (including painting, sculpture, and photography) interact: through the depiction of such earth sciences phenomena as mountains or storms, through the use of actual geologic illustrations or even maps as forms of artwork, through the application of geologic materials in art (most notably, marble in sculpture), and

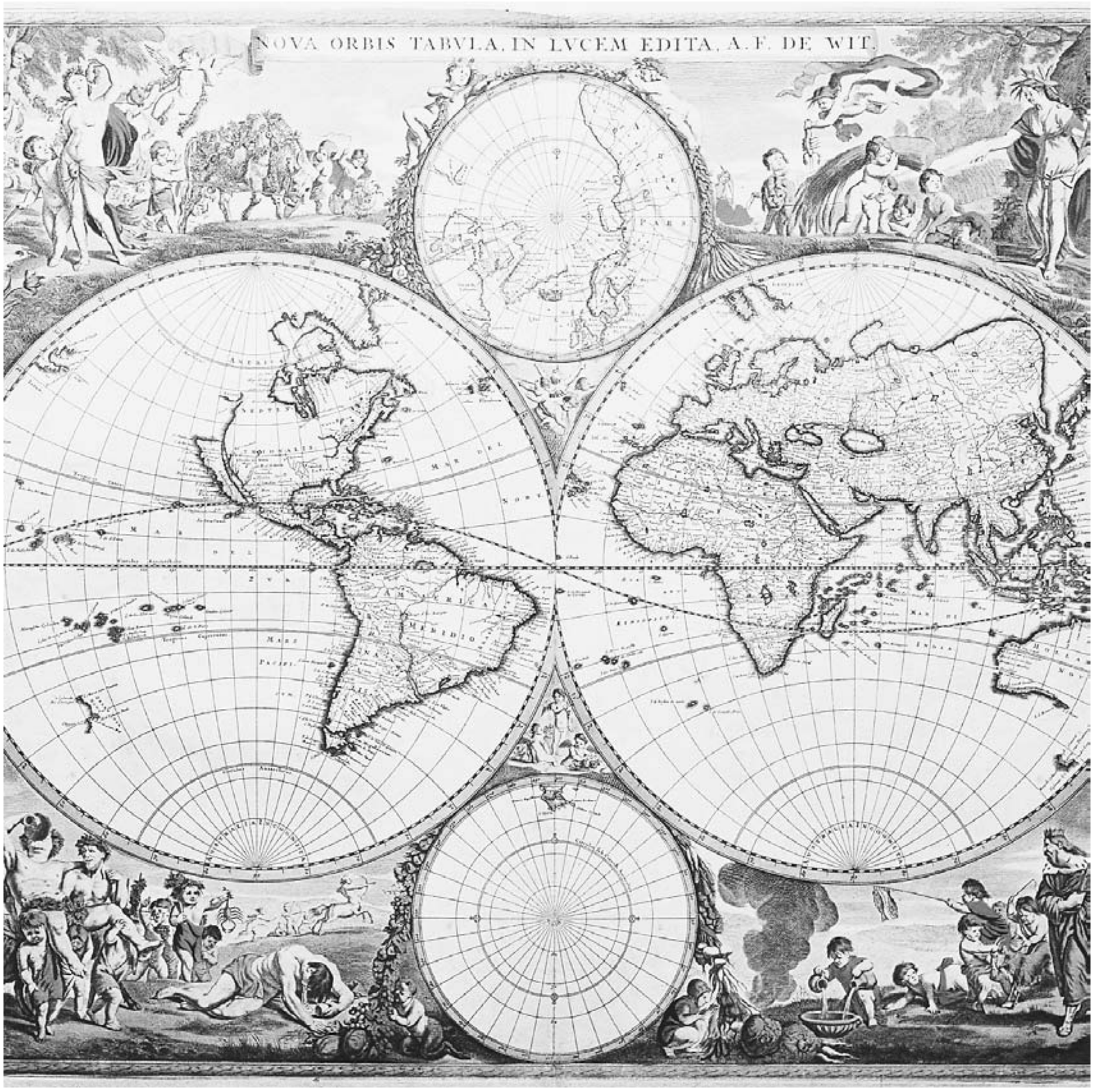
While stone is a strong building material in terms of compression, it is weak in terms of tension. The arch owes its strength to the principle of weight distribution, which overcomes this shortcoming of stone. Indeed, the Roman Coliseum has stood for more than two thousand years. *(© John Moss/Photo Researchers. Reproduced by permission.)*

through the employment of geology to investigate aspects of art objects (for instance, determining the origins of materials in ancient sculpture).

In the first category, visual depictions of geologic phenomena, Scott mentions works by unknown artists of various premodern civilizations (in particular, China and Japan) as well as by more recent artists whose names are hardly household words. On the other hand, some extremely well known figures produced notable works related to geoscience and the earth sciences. For example, the Italian artist and scientist Leonardo da Vinci (1452–1519), who happened to be one of the fathers of geology (see Studying Earth), painted many canvases in which he portrayed landscapes with a scientist's eye.

Another noteworthy example of earth sciences artwork and illustration is *The Great Piece of Turf* (1503), by Leonardo's distinguished contemporary the German painter and engraver Albrecht Dürer (1471–1528). A life-size depiction of grasses and dandelions, *Turf* belongs within the realm of earth sciences or even biological sciences rather than geoscience, yet it is significant as a historical milestone for all natural sciences.

In creating this work, Dürer consciously departed from the tradition, still strong even in



Scientific illustration became popular between 1500 and 1700, bridging the boundary between earth science and art. This map of the world, surrounded by allegorical scenes depicting the rewards and pitfalls of exploration, dates to 1689. *(© G. Bernard/Photo Researchers. Reproduced by permission.)*

the Renaissance, of representing "important" subjects, such as those of the Bible and classical mythology or history. By contrast, Dürer chose a simple scene such as one might find at the edge of any pond, yet his painting had a tremendous artistic and scientific impact. He set a new tone of naturalism in the arts and established a standard for representing nature as it is rather than in the idealized version of the artist's imagination.

As a result of Dürer's efforts, the period between about 1500 and 1700 saw the appearance of botanical illustrations whose quality far exceeded that of all previous offerings. Thus, he started a movement that spread throughout the world of scientific illustrations in general. Later, such geologists as England's William Smith (1769–1839) would produce maps that are rightly regarded as works of art in their own right (see Measuring and Mapping Earth).

Sometimes geologic phenomena have themselves become the basis for works of art, as Scott points out, observing that the modern American artist James Turrell once "set out to modify an extinct volcano, the Roden Crater [in northern Arizona], by excavating chambers and a tunnel to provide a visual experience of varying spatial relationships, the effects of light, and the perception of the sky." Elsewhere in the *Oxford Companion,* other writers show how evidence of a geoscientific influence has appeared in other arts and media, including music.

MUSIC. In "Music and the Earth Sciences," D. L. Dineley and B. Wilcock offer a fascinating overview of natural formations or materials that have their own musical qualities: for example, the "singing sands" of the Arabian peninsula and other regions, which produce musical tones when millions of grains are rubbed together by winds. The authors also discuss the effect of geologic phenomena on the sound and production of music—for instance, the acoustic qualities of music played in an auditorium built of stone.

Then there is the subject of musical compositions inspired by geoscientific or earth sciences phenomena. Among them are *The Hebrides; or, Fingal's Cave* by the German composer Felix Mendelssohn (1809–1847) as well as one the authors do not mention: *The Planets,* presented in 1918 by the German composer Gustav Holst (1874–1934). One also might list popular songs that refer to such phenomena, including "The White Cliffs of Dover." Written by Walter Kent and Nat Burton in 1941, the song epitomizes the longing for peace in a world torn by war. The cliffs themselves, which guard the eastern approaches of Britain, sometimes are referred to incorrectly as "chalk," though they are made of gypsum.

Ironically, rock music has few significant songs that refer to rocks. Usually the language is metaphoric, as was the case with the Bob Seger song discussed earlier. Hence, we have the name of the rock group Rolling Stones (with its implicit reference to the proverbial saying mentioned earlier) as well as the title to one of their earliest hits, "Heart of Stone." Jim Morrison's lyrics for the Doors include several references to the ground and things underneath it, including a gold mine in "The End." Coal mines have appeared in more than one song: "Working in the Coal Mine" was a hit for Lee Dorsey in the 1960s and was performed anew by the group Devo in 1981—not long after the Police song "Canary in a Coal Mine" appeared.

FILM. More significantly, the year 1981 marked the release of *Raiders of the Lost Ark,* a film cited as a major turning point by Ted Nield in the *Oxford Companion*'s "Geoscience in the Media" entry. The film is not about a geoscientist but an archaeologist, Indiana Jones (played by the actor Harrison Ford); however, the character of Jones is based on an American paleontologist, Roy Chapman Andrews (1884–1960). Earlier movies, Nield observes, had portrayed the typical scientist as an "egghead . . . an arrogant, unworldly, megalomaniac obsessive . . . But with Indiana Jones we saw the beginning of a reaction. Increasing audience sophistication is part of the reason."

Nield goes on to discuss the movie *Jurassic Park* (1993), which features three scientists, all of whom receive positive treatment. The actor Sam Neill, as a paleontologist, is described as "dedicated—perhaps a bit too educated—but also intuitive, a superb communicator, and above all, knowledgeable about dinosaurs." Laura Dern, playing a paleobiologist, is "strong-willed, independent, feminist, and sexy," while Jeff Goldblum's mathematician is "weird, roguish, and cool." Sparking a widespread interest in dinosaurs and paleontology, the film (a major box-office hit directed by Steven Spielberg) helped advance the cause of the geosciences.

The positive trend in movie portrayals of geoscientists, Nield states, continued in *Dante's Peak* (1997), in which even the casting of the ultra-handsome actor Pierce Brosnan as a geologist says a great deal about changing perceptions of scientists. Noting that audiences had come to differentiate between science and the misapplication thereof, Nield observes that "The heat seems to have come off those who are merely curious about Nature's workings." Additionally, "by being associated with the open air and fieldwork, [geoscientists] can take on some of the clichéd but healthy characteristics usually associated on film with oilmen and lumberjacks."

In an entirely different category is another fascinating example of geoscience in film, Australian director Peter Weir's *Picnic at Hanging Rock* (1975). Weir, who went on to make such well-known films as *The Year of Living Dangerously* (1982), *Witness* (1985), and *Dead Poets' Society* (1989), established his reputation—and that of Australian cinema in general—with *Picnic,* which concerns the disappearance of a group of schoolgirls and their teacher on Valentine's Day, 1900. The story itself is fictional, though it seems otherwise (*Picnic* later inspired *The Blair Witch Project,* which also presents fiction as fact); however, the rock in the title is very much a real place. In the film, Hanging Rock is by far the most striking character, a brooding presence whose foreboding features serve as a reminder of Earth's vastness and great age in the face of human insignificance.

The Work of Geoscientists

The work of the geoscientist indeed is associated with the open air to a much greater degree than that of the physicist or chemist; on the other hand, a geoscientist might very well work indoors, for instance, as a teacher. Prospective geoscientists who subscribe to a worldview of environmental utopianism can get a job "saving the world"—perhaps even working for starvation wages, so as to heighten the nobility of the undertaking. On the other hand, a pragmatist can go to work for an "evil" oil company and make a good living. The point is that there is a little of something for everyone in the world of geoscience.

Geoscientists may work for educational institutions, governments, or private enterprise. They may be involved in the search for energy resources, such as coal or oil (or even uranium for nuclear power), or they may be put to work searching for valuable and precious metals ranging from iron to gold. They even may be employed in the mining of diamonds or other precious gems in South Africa, Russia, or other locales. Other, perhaps less glamorous but no less important resources for which geoscientists in various roles search are water as well as rocks, clay, and minerals for building.

The majority of employed geoscientists work for industry but not always in the capacity of resource extraction. Some are involved in environmental issues; indeed, environmental geology—the application of geologic techniques to analyze, monitor, and control the environmental impact of natural and human phenomena—is a growing field. Among the areas of concern for environmental geologists are water management, waste disposal, and land-use planning.

ENVIRONMENTAL AND URBAN GEOLOGY. Many environmental geologists, as one might expect, are employed by governments. They may be involved in soil studies before the commencement of a building project, in analyzing the necessary thickness and materials for a particular stretch of road, or in designing and establishing specifications for a landfill. Many such concerns come into play when large populations gather together. In fact, a growing area of specialization in environmental geology is urban geology.

Urban geology can be defined as the application of geologic techniques to the study of the built environment. (The latter term is architectural and engineering jargon for any physical or geographic area containing human construction.) At first, "urban geology" might almost seem like an oxymoron, since the term geology usually calls to mind vast, unpopulated mountain ranges and rock formations—perhaps in South Dakota or Wyoming. In fact, geology is a major factor in the development of cities. Most are defined by their geomorphology: the hills of Athens and Rome, the mountains above Los Angeles, or the harbors of New York and other major ports, for instance.

Most cities have natural barriers to growth, and this is precisely because geomorphology originally dictated the location at which the city was established. A rare exception is Atlanta, Georgia, which grew around the point where several rail lines met. (In the 1840s, when it was established, it bore the name Terminus, a reference to the fact that it lay at the end of the rail line.) Bounded by no ocean, significant rivers, mountains, or other natural barriers, such as deserts, Atlanta began a period of explosive growth in the latter part of the twentieth century and has never stopped growing. Today Atlanta is a textbook example of urban sprawl: lacking a vital city center, it is a settlement of some four million people spread over an area much larger than Rhode Island, with no end to growth in sight.

Los Angeles often is cited as a case of urban sprawl, but its problems are quite different: it is rife with geomorphologic barriers, including oceans, mountains, and desert. The result is increasing growth within a limited area, resulting in heightened stress on existing resources. These are some of the issues confronted by urban geologists. Another example is the problem of determining the strength of bedrock, which dictates the viability of tall buildings. Urban geologists also are concerned with such issues as underground facilities for transportation, infrastructure, and even usable workspace—one possible solution to the problem of urban sprawl.

GEOARCHAEOLOGY AND RELATED FIELDS. At the opposite extreme, in many ways, from urban geology is geoarchae-

A gold mine in Zimbabwe. Geoscientists work for educational institutions, governments, and private enterprise in such fields as resource extraction, environmental studies and management, and even archaeology and criminology. *(© Peter Bowater/Photo Researchers. Reproduced by permission.)*

ology, or the application of geologic analysis to archaeology and related fields. Whereas urban geology is concerned with the here and now, geoarchaeology—like the larger field of historical geology—addresses the past. And whereas urban geologists are most likely to be employed by governments, geoarchaeologists and those in similar areas are typically on the payroll of universities.

In a different sense, geoarchaeology also contrasts with archaeological geology, which is the study of archaeological sites for data relevant to the geosciences; thus, archaeological geology stands the approach of geoarchaeology on its head. An example of a study in archaeological geology can be found in the work conducted around the Roman ruins at Hierapolis in what is now Turkey. There, investigation of walls and gutters reveals the fact that the city was sitting astride an earthquake fault zone—a fact unknown to its residents, except when they experienced seismic tremors.

By contrast, an example of geoarchaeology in action would be establishing an explanation for how people came to the Americas from Siberia near the end of the last ice age—by crossing a land bridge that existed at that time. Another example of geoarchaeology would be the realm of ecclesiastical geology, which involves the study of old church masonry walls with the purpose of identifying areas from which rocks, bricks, and other materials were derived. Studies of medieval churches in England, for instance, show varieties of rock from sometimes unexpected locations, often placed alongside bricks taken from older Roman structures.

From the explanation and examples given here, it may be a bit hard to discern the difference between geoarchaeology and archaeological geology. Certainly there is a great deal of overlap, and in practice the difference comes down to a question of who is leading the fieldwork—a geologist or an archaeologist. In any case, both realms are concerned with the relatively recent human past, as opposed to the vast stretches of time that are the domain of historical geology (see Geologic Time).

FORENSIC GEOLOGY. On October 7, 2001, the United States launched air strikes against Afghanistan in retaliation for the refusal of that country's Taliban regime to surrender Osama bin Laden, the suspected mastermind of the World Trade Center bombing on September 11. On the same day, bin Laden's al-Qaeda ter-

KEY TERMS

ATMOSPHERIC SCIENCES: A major division of the earth sciences, distinguished from geoscience and the hydrologic sciences by its concentration on atmospheric phenomena. Among the atmospheric sciences are meteorology and climatology.

BIOSPHERE: A combination of all living things on Earth—plants, mammals, birds, reptiles, amphibians, aquatic life, insects, viruses, single-cell organisms, and so on—as well as all formerly living things that have not yet decomposed.

EARTH SCIENCES: The entire range of scientific disciplines focused on the study of Earth, including not only geoscience but also the atmospheric and hydrologic sciences.

ENVIRONMENTAL GEOLOGY: A field of geology involved in the application of geologic techniques to analyze, monitor, and control environmental impact of both natural and human phenomena.

GEOCHEMISTRY: A branch of the earth sciences, combining aspects of geology and chemistry, that is concerned with the chemical properties and processes of Earth—in particular, the abundance and interaction of chemical elements and their isotopes.

GEOLOGY: The study of the solid earth, in particular, its rocks, minerals, fossils, and land formations.

GEOMORPHOLOGY: An area of physical geology concerned with the study of landforms, with the forces and processes that have shaped them, and with the description and classification of various physical features on Earth.

GEOPHYSICS: A branch of the earth sciences that combines aspects of geology and physics. Geophysics addresses the planet's physical processes as well as its gravitational, magnetic, and electric properties and the means by which energy is transmitted through its interior.

rorist organization released a videotape of their leader delivering a diatribe against the United States. Naturally, military and law-enforcement agencies involved in the hunt for bin Laden took an interest in the tape, and some specialists sought clues in an unexpected place: the rocks behind bin Laden, featured prominently in the tape.

Although the efforts to trace bin Laden's location by the rock formations in the area were not successful, the underlying premise—that geographic regions have their own specific types and patterns of rock—was both a fascinating and a plausible one. This was just another example of a specialty known as forensic geology, or the use of geologic and other geoscientific data in solving crimes. Forensic geology has it origins around the beginning of the twentieth century, but some historians cite Sherlock Holmes, the master sleuth created by the English physician and writer Sir Arthur Conan Doyle (1859–1930), as an early practitioner.

In *The Sign of Four,* for instance, Holmes uses geologic data to ascertain that Watson has been to the Wigmore Street Post Office: "Observation tells me that you have a little reddish mould adhering to your instep," he explains. "Just opposite the Wigmore Street Office they have taken up the pavement and thrown up some earth, which lies in such a way that it is difficult to avoid treading in it in entering. The earth is of this peculiar reddish tint which is found, as far as I know, nowhere else in the neighbourhood."

The true founder of forensic geology was probably the Austrian jurist and pioneer in

KEY TERMS CONTINUED

GEOSCIENCE: The geologic sciences (geology, geochemistry, geophysics, and related disciplines), as opposed to other earth sciences—that is, atmospheric sciences, such as meteorology, and hydrologic sciences, such as oceanography.

GEOSPHERE: The upper part of Earth's continental crust, or that portion of the solid earth on which human beings live and which provides them with most of their food and natural resources.

HISTORICAL GEOLOGY: The study of Earth's physical history. Historical geology is one of two principal branches of geology, the other being physical geology.

HYDROLOGIC SCIENCES: Areas of the earth sciences concerned with the study of the hydrosphere. Among these areas are hydrology, glaciology, and oceanography.

HYDROSPHERE: The entirety of Earth's water, excluding water vapor in the atmosphere but including all oceans, lakes, streams, groundwater, snow, and ice.

ORGANIC: At one time chemists used the term *organic* only in reference to living things. Now the word is applied to most compounds containing carbon, with the exception of carbonates (which are minerals) and oxides, such as carbon dioxide.

PHYSICAL GEOLOGY: The study of the material components of Earth and of the forces that have shaped the planet. Physical geology is one of two principal branches of geology, the other being historical geology.

PHYSICAL SCIENCES: Astronomy, physics, chemistry, and the earth sciences.

SEDIMENT: Material deposited at or near Earth's surface from a number of sources, most notably preexisting rock. Soil is derived from sediment, particularly the mixture of rock fragments and organic material.

criminology Hans Gross (1847–1915), whose *Handbuch für Untersuchungsrichter* (Handbook for examining magistrates, 1898) was a pivotal work in the field. "Dirt on shoes," wrote Gross, "can often tell us more about where the wearer of those shoes has last been than toilsome inquiries." Near the turn of the nineteenth century, Germany's Georg Popp, who operated a forensic laboratory in Frankfurt, used the new science effectively in two cases.

The first of these cases involved the murder of a woman named Eva Disch in October 1904. Among the items found at the murder scene was a dirty handkerchief containing traces of coal, snuff, and hornblende, a mineral. Popp matched the handkerchief with a suspect who worked at two locations that used a great deal of hornblende. In addition, the suspect's pants cuffs bore soil both from the murder scene and the victim's house.

Four years later, in investigating the murder of Margaethe Filbert in Bavaria, Popp ascertained that the soil at the crime scene was characterized by red quartz and red clay rich in iron. By contrast, the chief suspect had a farm whose fields were notable for their porphyry, milky quartz, and mica content. As it turned out, the suspect's shoes bore traces of quartz and red clay rather than those other minerals, even though he claimed he had been working in his fields when the crime occurred.

WHERE TO LEARN MORE

Career Information for Geology Majors (Web site). <http://www.uakron.edu/ascareer/Geology.html>.

Careers in the Geosciences (Web site). <http://www.agiweb.org/agi/careers.html>.

Geoarchaeology (Web site). <http://www.geoarchaeology.com/geoarc.htm>.

Geoarchaeology and Related Subjects (Web site). <http://archaeology.about.com/cs/earthsciences/>.

Geology in the City (Web site). <http://www.fortunecity.com/greenfield/ecolodge/60/cityhome.html>.

Hancock, Paul L., and Brian J. Skinner, eds. *The Oxford Companion to the Earth.* New York: Oxford University Press, 2000.

Murray, Raymond C. *Devil in the Details: The Science of Forensic Geology* (Web site). <http://www.forensicgeology.net/science.htm>.

Sarah Andrews Forensic Geology Pages (Web site). <http://www.sonoma.edu/Geology/andrews.htm>.

Sevil Atasoy's Links of Forensic Geology (Web site). <http://abone.turk.net/atasoy/forensicgeology.html>.

Urban Geology of the National Capital (Web site). <http://sts.gsc.nrcan.gc.ca/page1/urban/urb.htm>.

What Can I Do with a Degree in Geology? (Web site). <http://www.hartwick.edu/geology/alumni/careers.html>.

EARTH SYSTEMS

CONCEPT

A system is any set of interactions set apart from the rest of the universe for the purposes of study, observation, and measurement. Theoretically, a system is isolated from its environment, but this is an artificial construct, since nothing is ever fully isolated. Earth is largely a closed system, meaning that it exchanges very little matter with its external environment in space, but the same is not true of the systems within the planet—geosphere, hydrosphere, biosphere, and atmosphere—which interact to such a degree that they are virtually inseparable. Together these systems constitute an intricate balance, a complex series of interrelations in which events in one sector exert a profound impact on conditions in another.

HOW IT WORKS

Systems

An isolated system is one so completely sealed off from its environment that neither matter nor energy passes through its boundaries. This is an imaginary construct, however, an idea rather than a reality, because it is impossible to create a situation in which no energy is exchanged between the system and the environment. Under the right conditions it is perhaps conceivable that matter could be sealed out so completely that not even an atom could pass through a barrier, but some transfer of energy is inevitable. The reason is that electromagnetic energy, such as that emitted by the Sun, requires no material medium in which to travel.

In contrast to an isolated system is a closed system, of which Earth is an approximation. Despite its name, a closed system permits the exchange of energy with the environment but does not allow matter to pass back and forth between the external environment and the system. Thus, Earth absorbs electromagnetic energy, radiated from the Sun, yet very little matter enters or departs Earth's system. Note that Earth is an *approximation* of a closed system: actually, some matter does pass from space into the atmosphere and vice versa. The planet loses traces of hydrogen in the extremities of its upper atmosphere, while meteorites and other forms of matter from space may reach Earth's surface.

Earth more closely resembles a closed system than it does an open one—that is, a system that allows the full and free exchange of both matter and energy with its environment. The human circulatory system is an example of an open system, as are the various "spheres" of Earth (geosphere, hydrosphere, biosphere, and atmosphere) discussed later. Whereas an isolated system is imaginary in the sense that it does not exist, sometimes a different feat of imagination is required to visualize an open system. It is intricately tied to its environment, and therefore the concept of an open system as a separate entity sometimes requires some imagination.

Using Systems in Science

To gain perspective on the use of systems in science as well as the necessity of mentally separating an open system from its environment, consider how these ideas are used in formulating problems and illustrating scientific principles. For example, to illustrate the principle of potential and kinetic energy in physics, teachers often use the example of a baseball dropping from a great height (say, the top of a building) to the ground.

At the top of the building, the ball's potential energy, or the energy it possesses by virtue of its position, is at a maximum, while its kinetic energy (the energy it possesses by virtue of its motion) is equal to zero. Once it is dropped, its potential energy begins to decrease, and its kinetic energy to increase. Halfway through the ball's descent to the ground, its potential and kinetic energy will be equal. As it continues to fall, the potential energy keeps decreasing while the kinetic energy increases until, in the instant it strikes the ground, kinetic energy is at a maximum and potential energy equals zero.

KEEPING OUT IRRELEVANT DETAILS. What has been described here is a system. The ball itself has neither potential nor kinetic energy; rather, energy is in the system, which involves the ball, the height through which it is dropped, and the point at which it comes to a stop. Furthermore, because this system is concerned with potential and kinetic energy only in very simple terms, we have mentally separated it from its environment, treating it as though it were closed or even isolated, though in reality it would more likely be an open system.

In the real world, a baseball dropping off the top of a building and hitting the ground could be affected by such conditions as prevailing winds. These possibilities, however, are not important for the purposes of illustrating potential and kinetic energy, and even if they were, they could be incorporated into the larger energy system.

THE "MAGIC" OF A SYSTEM. Since kinetic energy and potential energy are inversely related, the potential energy at the top of the building will always equal the kinetic energy at the point of maximum speed, just before impact. This is true whether the ball is dropped from 10 ft. (3 m) or 1,000 ft. (305 m). It may seem almost magical that the sum of potential and kinetic energy is always the same or that the two values are perfectly inverse. In fact, there is nothing magical here: the system has a certain total energy, and this does not change, though the distribution of that energy can and does vary.

Suppose one had a money jar known to contain $20. If one reaches in and grasps a five-dollar bill, two one-dollar bills, three quarters, a dime, and two nickels ($7.95), there must be $12.05 left in the jar. There is nothing magical in this; rather, what has been illustrated is the physical principle of conservation. In physics and other sciences, "to conserve" something means "to result in no net loss of" that particular component. It is possible that within a given system, the component may change form or position, but as long as the net value of the component remains the same, it has been conserved. Thus, the total energy is conserved in the situation involving the baseball, and the total amount of money is conserved in the money-jar.

APPLYING THE SYSTEM PRINCIPLE TO EARTH

In the baseball illustration, the distribution between types of energy varies, but the total amount is always the same. Likewise in the money-jar illustration, the total amount of money remains fixed even though the distribution according to various denominations may vary. The same is true of Earth, though here it is the total amount of matter. This includes valuable resources, among them materials that can be mined to produce energy—for instance, fossil fuels such as coal or petroleum—as well as waste products. Because Earth is a closed system, there are no additional resources, nor is there any dumping ground other than the one beneath our feet. Thus, the situation calls for prudence both in the use of the planet's material wealth and in the processing of materials that will leave a by-product of waste.

The fact that a closed system is by definition finite leads to the principle that the relationships between its constituent parts are likewise finite, and therefore changes in one part of the system are liable to produce effects in another part. Conditions in the baseball or money-jar illustrations are so simple that it is easy to predict the effect of a change. For instance, if we substitute a basketball for a baseball, this will change the total energy, because the latter is a function of the ball's mass. If the denominations making up the $20 in the money jar are replaced with a collection of two-dollar bills and dimes, this will make it impossible to reach in and pull out an odd-numbered value in dollars or cents.

What about the changes that result when one aspect of Earth's system is altered? In some cases, it is easy to guess; in others, the interactions are so complex that prediction requires sophisticated mathematical models. It is perhaps no accident that chaos theory was developed by a meteorologist, the American Edward Lorenz (1917–). Chaos theory, the study of complex systems that appear to follow no orderly laws, involves the analysis of phenomena that appear connected by something than an ordinary cause and effect relationship. The classic example of this is the "butterfly effect, " the idea that a butterfly beating its wings in China can change the weather in New York City. This, of course, is a far-fetched scenario, but sometimes changes in one sector of Earth's system can yield amazing consequences in an entirely different part.

THE FOUR "SPHERES"

The systems approach is relatively new to the earth sciences, themselves a group of disciplines whose diversity reflects the breadth of possible approaches to studying Earth (see Studying Earth). At one time, earth scientists tended to investigate specific aspects of Earth without recognizing the ways in which these aspects connect with one another; today, by contrast, the paradigm of the earth sciences favors an approach that incorporates the larger background.

Given the complexities of Earth itself, as well as the earth sciences, it is helpful to apply a schema (that is, an organizational system) for dividing larger concepts and entities into smaller ones. For this reason, earth scientists tend to view Earth in terms of four interconnected "spheres. " One of these terms, atmosphere, is a familiar one, while the other three (geosphere, hydrosphere, and biosphere) may sound at first like mere scientific jargon.

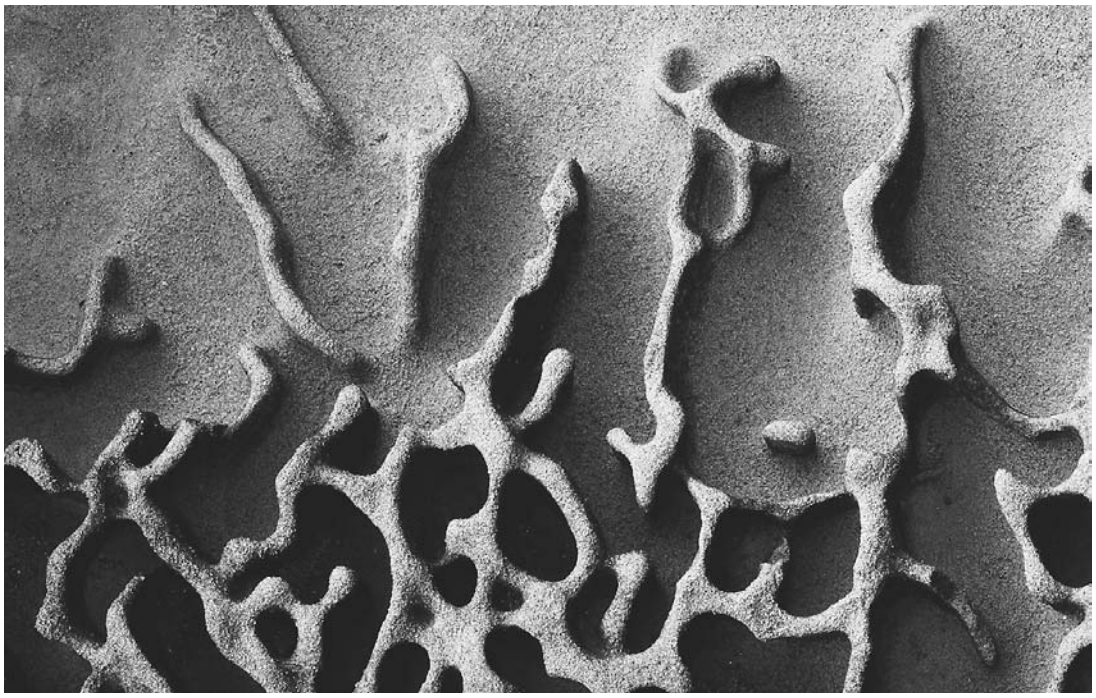
UNDERSTANDING THE SPHERES. In fact, each sphere represents a sector of existence on the planet that is at once clearly defined and virtually inseparable from the others. Each is an open system within the closed system of Earth, and overlap is inevitable. For example, the seeds of a plant (biosphere) are placed in the ground (geosphere), from which they receive nutrients for growth. In order to sustain life, they receive water (hydrosphere) and carbon dioxide (atmosphere). Nor are they merely receiving: they also give back oxygen to the atmosphere, and by providing nutrition to an animal, they contribute to the biosphere.

Each of the spheres, or Earth systems, is treated in various essays within this book. These essays examine these subsystems of the larger Earth system in much greater depth; what follows, by contrast, is the most cursory of introductions. It should be noted also that while these four subsystems constitute the entirety of Earth as humans know and experience it, they are only a small part of the planet's entire mass. The majority of that mass lies below the geosphere, in the region of the mantle and core.

A CURIOUS AND INSTRUCTIVE POINT. As a passing curiosity, it is interesting to note that modern scientists have identified four subsystems and given them the name spheres. As discussed in the essay Earth, Science, and Nonscience, the ancient Greeks were inclined to divide natural phenomena into fours, a practice that reached its fullest expression in the model of the universe developed by the Greek philosopher Aristotle (384–322 B.C.) He even depicted the physical world as a set of spheres and suggested that the heaviest material would sink to the interior of Earth while the lightest would rise to the highest points.

These points of continuity with ancient science are notable because almost everything about Aristotle's system was wrong, and, indeed, the differences between his model of the physical world and the modern one are instructive. There are four spheres in the modern earth sciences because these four happen to be useful ways of discussing the larger Earth system—not, as in the case of the Greeks, because the number four represents spiritual perfection. Furthermore, scientists understand these spheres to be artificial constructs, at least to some extent, rather than a key to some deeper objective reality about existence, as the ancients would have supposed.

Nor are the spheres of the modern earth sciences literally spheres, as Aristotle's concentric orbits of the planets around Earth were. If anything, the use of the term sphere represents a holdover from the Greek way of viewing the material world. Finally, unlike such ancient notions as the concept of the four elements, four qualities, or four humors, the idea of the four spheres is not simply the result of pure conjecture. Instead, the concept of these four interrelated systems came about by application of the scientific method and entered the vocabulary of earth scientists because the ideas involved clearly reflected and illustrated the realities of Earth processes.



SANDSTONE ERODED BY WAVES. *(© Stephen Parker/Photo Researchers. Reproduced by permission.)*

THE SPHERES IN BRIEF

The geosphere itself may be defined as the upper part of the planet's continental crust, the portion of the solid earth on which human beings live, which provides them with most of their food and natural resources. Even with the exclusion of the mantle and core, the solid earth portion of Earth's system is still by far the most massive. It is estimated that the continental and oceanic crust to a depth of about 1.24 mi. (2 km) weighs 6×10^{21} kg—about 13,300 billion billion pounds. The mass of the biosphere, by contrast, is about one millionth that figure. If the mass of all four spheres were combined, the geosphere would account for 81.57%, the hydrosphere 18.35%, the atmosphere 0.08%, and the biosphere a measly 0.00008%. (Of that last figure, incidentally, animal life—of which humans are, of course, a very small part—accounts for less than 2%.)

Not only is the geosphere the largest, it is also by far the oldest of the spheres. Its formation dates back about four billion years, or within about 0.5 billion years of the planet's formation. As Earth cooled after being formed from the gases surrounding the newborn Sun, its components began to separate according to density. The heaviest elements, such as iron and nickel, drifted toward the core, while silicon rose to the surface to form the geosphere.

ATMOSPHERE, HYDROSPHERE, AND BIOSPHERE. In that distant time Earth had an atmosphere in the sense that there was a blanket of gases surrounding the planet, but the atmospheric composition was quite different from today's mixture of nitrogen (78%), oxygen (21%), argon (0.93%), and other substances that include water vapor, carbon dioxide, ozone, and noble gases such as neon, which together comprise 0.07%. The atmosphere then consisted largely of carbon dioxide from Earth's interior as well as gases brought to Earth by comets. Elemental hydrogen and helium escaped the planet, and much of the carbon was deposited in what became known as carbonate rocks. What remained was a combination of hydrogen compounds, including methane, ammonia, nitrogen- and sulfur-rich compounds expelled by volcanoes, and (most important of all) H_2O, or water.

Simultaneous with these developments, the gases of Earth's atmosphere cooled and condensed, taking the form of rains that, over millions of years, collected in deep depressions on the planet's surface. This was the beginning of the oceans, the largest but far from the only component of Earth's hydrosphere, which consists of all the planet's water except for water vapor in the atmosphere. Thus, the hydrosphere includes not only saltwater but also lakes, streams, groundwater, snow, and ice.

Water, of course, is necessary to life, and it was only after its widespread appearance that the first life-forms appeared. This was the beginning of the biosphere, which consists of all living organisms as well as any formerly living material that has not yet decomposed. (Typically, following decomposition an organism becomes part of the geosphere.) Over millions of years, plants formed, and these plants gradually began producing oxygen, helping to create the atmosphere as it is known today—an example of interaction between the open systems that make up the larger Earth system.

REAL-LIFE APPLICATIONS

EARTH AS AN ORGANISM

Clearly, a great deal of interaction occurs between spheres and has continued to take place for a long time. Earth often is described as a living organism, a concept formalized in the 1970s by the English meteorologist James Lovelock (1919–) and the American biologist Lynn Margulis (1938–), who developed the Gaia hypothesis. Sometimes called the Gaian hypothesis, this principle is named after the Greek earth goddess, a prototype for "Mother Earth," and is based on the idea that Earth possesses homeostatic or self-regulating mechanisms that preserve life. (Lovelock's neighbor William Golding [1911–1993], author of *Lord of the Flies,* suggested the name to him.)

Though the Gaia hypothesis seems very modern and even a bit "New Age" (that is, relating to a late twentieth-century movement that incorporates such themes as concern for nature and spirituality), it has roots in the ideas of the great Scottish geologist James Hutton (1726–1797), who described Earth as a "superorganism." A forward-thinking person, Hutton maintained that physiology provides the model for the study of Earth systems. Out of Hutton's and, later, Lovelock's ideas ultimately grew the earth science specialty of geophysiology, an interdisciplinary approach incorporating aspects of geochemistry, biology, and other areas.

The Gaia hypothesis is far from universally accepted, however, and remains controversial. One reason is that it seems to contain a teleologic, or goal-oriented, explanation of physical behaviors that does not fully comport with the findings of science. An animal responds to external conditions in such a way as to preserve life, but this is because it has instinctive responses "hardwired" into its brain. Clearly, if the Earth is an "organism, " it is an organism in quite a different sense than an animal, since it does not make sense to describe Earth as having a "brain."

HOMEOSTASIS AND CYCLES

Nonetheless, Lovelock, Margulis, and other supporters of the Gaia hypothesis have pointed to a number of anomalies that have yet to be explained fully and for which the Gaia hypothesis offers one possible solution. For example, it would have taken only about 80 million years for the present levels of salt in Earth's oceans to have been deposited there from the geosphere; why, then, is the sea not many, many times more salty than it is? Could it be that Earth has somehow regulated the salinity levels in its own seas?

Earth's systems unquestionably display a homeostatic and cyclical behavior typical of living organisms. Just as the human body tends to correct any stresses imposed on it, Earth likewise seeks equilibrium. And just as blood, for instance, cycles through the body's circulatory system, so matter and energy move between various spheres in the course of completing certain cycles of the Earth system. These include the energy and hydrologic cycles; a number of biogeochemical cycles, such as the carbon and nitrogen cycles; and a rock cycle of erosion, weathering, and buildup. (Each of these systems is discussed in a separate essay, or as part of a separate essay, in this book.)

FEEDBACK. Though particulars of the Gaia hypothesis remain a matter of question, it is clear that Earth regulates these cycles and does so through a process of feedback and corrections. To appreciate the idea of feedback, consider a financial example. In the early 1990s, the U.S. Congress placed a steep tax on luxury boats, presumably with the aim of getting more money from wealthy taxpayers. The result, however, was exactly the opposite: boat owners sold their crafts, and many of those considering purchases cancelled their plans. Rather than redistributing

AN OIL-COVERED BIRD, VICTIM OF THE 1989 EXXON VALDEZ'S OIL SPILL IN PRINCE WILLIAM SOUND, ALASKA. *(AP/Wide World Photos. Reproduced by permission.)*

wealth from the rich to those less fortunate, the tax resulted in the government's actually getting *less* money from rich yacht owners.

Whereas Congress expected the rich to provide positive feedback by giving up more tax money, instead the yacht owners responded by acting against the tax—a phenomenon known as negative feedback. Feedback itself is the return of output to a system, such that it becomes input which then produces further output. Feedback that causes the system to move in a direction opposite that of the input is negative feedback, whereas positive feedback is that which causes the system to move in the same direction as the input. The luxury tax would have made perfect sense if the purpose had been to halt the production and purchase of expensive boats, in which case the output would have been deemed positive.

In the luxury-tax illustration, negative feedback is truly "negative" in the more common sense of the word, but this is not typically the case where nature in general or Earth systems in particular are concerned. In natural systems negative feedback serves as a healthy corrective and tends to stabilize a system. To use an example from physiology, if a person goes into a cold environment, the body responds by raising the internal temperature. Likewise, in chemical reactions the system tends to respond to any stress placed on it by reducing the impact of the stress, a concept known as Le Châtelier's principle after the French chemist Henry Le Châtelier (1850–36).

Positive feedback, on the other hand, is often far from "positive" and is sometimes described as a "vicious cycle." Suppose rainwater erodes a portion of a hillside, creating a gully. Assuming the rains continue, the opening of this channel for the water facilitates the introduction of more water and therefore further erosion of the hillside. Given enough time, the rain can wash a deep gash into the hill or even wash away the hill entirely.

FAR-REACHING CONSEQUENCES

Given the interconnectedness of systems on Earth, it is easy to see how changes in one part of the larger Earth system can have far-reaching impacts on another sector. For example, the devastating Alaska earthquake of March 1964 produced tsunamis felt as far away as Hawaii, while the Exxon *Valdez* oil spill that afflicted Alaska exactly 25 years later had an effect on the biosphere and hydrosphere over an enormous area.

El Niño is a familiar example of far-reaching consequences produced by changes in Earth sys-

KEY TERMS

ATMOSPHERE: A blanket of gases surrounding Earth and consisting of nitrogen (78%), oxygen (21%), argon (0.93%), and other substances that include water vapor, carbon dioxide, ozone, and noble gases such as neon, which together comprise 0.07%.

BIOSPHERE: A combination of all living things on Earth—plants, mammals, birds, reptiles, amphibians, aquatic life, insects, viruses, single-cell organisms, and so on—as well as all formerly living things that have not yet decomposed. Typically, after decomposing, a formerly living organism becomes part of the geosphere.

CLOSED SYSTEM: A system that permits the exchange of energy with its external environment but does not allow matter to pass between the environment and the system. Compare with isolated system, on the one hand, and open system, on the other.

CONSERVATION: In physics and other sciences, "to conserve" something means "to result in no net loss of" that particular component. It is possible that within a given system, the component may change form or position, but as long as the net value of the component remains the same, it has been conserved.

ELECTROMAGNETIC ENERGY: A form of energy with electric and magnetic components, which travels in waves and, depending on the frequency and energy level, can take the form of long-wave and short-wave radio; microwaves; infrared, visible, and ultraviolet light; x rays; and gamma rays.

ENVIRONMENT: In discussing systems, the term *environment* refers to the surroundings—everything external to and separate from the system.

FEEDBACK: The return of output to a system, such that the output becomes input that produces further output. Feedback that causes the system to move in a direction opposite to that of the input is negative feedback, whereas positive feedback is that which causes the system to move in the same direction as the input.

GAIA HYPOTHESIS: The concept, introduced in the 1970s, that Earth behaves much like a living organism, possessing self-regulating mechanisms that preserve life. Sometimes called the Gaian hypothesis, it is named after Gaia, the Greek goddess of the earth.

GEOSPHERE: The upper part of Earth's continental crust, or that portion of the solid earth on which human beings live and which provides them with most of their food and natural resources.

HOMEOSTASIS: A tendency toward equilibrium.

HOMEOSTATIC: The quality of being self-regulating.

HYDROSPHERE: The entirety of Earth's water, excluding water vapor in the atmosphere but including all oceans, lakes, streams, groundwater, snow, and ice.

ISOLATED SYSTEM: A system that is so fully separated from the rest of the universe that it exchanges neither matter nor energy with its environment. This is an imaginary construct, since full isolation is impossible.

OPEN SYSTEM: A system that allows complete, or near-complete, exchange of matter and energy with its environment.

KEY TERMS CONTINUED

SCIENTIFIC METHOD: A set of principles and procedures for systematic study that includes observation; the formation of hypotheses, theories, and laws; and continual testing and reexamination.

SYSTEM: Any set of interactions that can be set apart mentally from the rest of the universe for the purposes of study, observation, and measurement.

TSUNAMI: A tidal wave produced by an earthquake or volcanic eruption. The term comes from the Japanese words for "harbor" and "wave."

tems. Spanish for "child" (because it typically occurs around Christmastime), El Niño begins on the western coast of South America. There, every few years, trade winds slacken, allowing the wind from the west to push warm surface water eastward. Lacking vital nutrients, this warm water brings about a decline in the local marine life. It also causes heavy rains and storms.

IMPACT OF EL NIÑO AROUND THE WORLD. To the extent described, El Niño is largely a local phenomenon. But it can affect the jet streams, or high-level winds, that push storms across the Western Hemisphere. This can result in milder weather for western Canada or the northern United States, as the winds push more severe storms into Alaska, but it also can bring about heavy rains in the Gulf of Mexico region. Nor are its effects limited to the Western Hemisphere. El Niño has been known to alter the pattern of monsoons, or rainy seasons, in India, Southeast Asia, and parts of Africa, thus producing crop failures that affect millions of people.

Aside from the indirect effects, such as the famines in the Eastern Hemisphere, the direct effects of the El Niño phenomenon can be devastating. The El Niño of 1982–83, which affected the United States, the Caribbean, western South America, Africa, and Australia, claimed some 2,000 lives and cost about $13 billion in property damage. It returned with a vengeance 15 years later, in 1997–98, killing more than 2,100 people and destroying $33 billion worth of property.

Years Without Summer

Whereas El Niño is an example of a disturbance in the hydrosphere that affects the atmosphere and ultimately the biosphere, an even more terrifying phenomenon can begin with an eruption in the geosphere, which spreads to the atmosphere and then the hydrosphere and biosphere. This phenomenon might be called "years without summer"; an example occurred in 1815–16.

In June of 1816 snow fell in New England, and throughout July and August temperatures hovered close to freezing. Frosts hit in September, and New Englanders braced themselves for an uncommonly cold winter, as that of 1816–17 turned out to be. It must have seemed as though the world were coming to an end, yet the summer of 1817 proved to be a normal one. The cause behind this year without summer in 1816 lay in what is now Indonesia, and it began a year earlier.

In 1815, Mount Tambora to the east of Java had erupted, pouring so much volcanic ash into the sky that it served as a curtain against the Sun's rays, causing a brutally cold summer in New England the following year. An eruption of Mount Katmai in Alaska in 1912 produced far-reaching effects, including some lowering of temperatures, but its impact was nothing like that of Tambora. Nor did the 1980 Mount Saint Helens eruption in Washington State prove nearly as potent in the long run as the eruption of Tambora did (though it produced a devastating immediate impact).

THE CATACLYSM OF A.D. 535. Even the eruption of Mount Tambora may have been overshadowed by another, similar event, known simply as the catastrophe, or cataclysm, of A.D. 535. In the late twentieth century, the British dendrochronologist Mike Baillie discovered a pattern of severely curtailed growth in tree rings dating to the period A.D. 535–541. More or less simultaneous with Baillie's work was that of the amateur archaeologist David Keys, who found a number of historical texts by Byzantine, Chinese, and Anglo-Saxon scholars of the era, all suggesting that something cataclysmic had happened in A.D. 535. For example, the Byzantine historian Procopius (d. 565) wrote, "The sun gave forth its light without brightness ... for the whole year."

Some geologists have maintained that the cataclysm resulted from the eruption of another Indonesian volcano, the infamous Krakatau, which had a devastating eruption in 1883 and which could have produced enough dust to cause an artificial winter. Whatever the cause, the cataclysm had an enormous impact that redounds from that time perhaps up to the present. The temperature drop may have sparked a chain of events, beginning in southern Africa, that ultimately brought a plague to the Byzantine Empire, forcing Justinian I (r. A.D. 483–565) to halt his attempted reconquest of western Europe. At the same time, the cataclysm may have been responsible for food shortages in central Asia, which spawned a new wave of European invasions, this time led by the Avars.

The result was that the fate of Europe was sealed. For a few years it had seemed that Justinian could reconquer Italy, thus reuniting the Roman Empire, whose western portion had ceased to exist in A.D. 476. Forced to give up their reconquest, with the Avars and others overrunning Europe while the plague swept through Greece, the Byzantines turned their attention to affairs at home and increasingly shut themselves off from western Europe. Thus the Dark Ages, the split between Catholicism and Eastern Orthodoxy, the Crusades—even the Cold War, which reflected the old east-west split in Europe—may have been the results of a volcano on the other side of the world.

WHERE TO LEARN MORE

Cox, Reg, and Neil Morris. *The Natural World.* Philadelphia: Chelsea House, 2000.

Earth's Energy Budget (Web site). <http://radar.metr.ou.edu/OK1/meteorology/EnergyBudget.html.>.

Farndon, John. *Dictionary of the Earth.* New York: Dorling Kindersley, 1994.

The Gaia Hypothesis (Web site). <http://www.magna.com/au/~prfbrown/gaia_jim.html>.

Geophysiology Online (Web site). <http://www.webtorre.com/geo/GeoPhysOnline.asp>.

Hancock, Paul L., and Brian J. Skinner. *The Oxford Companion to the Earth.* New York: Oxford University Press, 2000.

Knapp, Brian J. *Earth Science: Discovering the Secrets of the Earth.* Illus. David Woodroffe and Julian Baker. Danbury, CT: Grolier Educational, 2000.

Kump, Lee R., James F. Kasting, and Robert G. Crane. *The Earth System.* Upper Saddle River, NJ: Prentice Hall, 2000.

Lovelock, J. E. *Gaia: A New Look at Life on Earth.* New York: Oxford University Press, 2000.

Skinner, Brian J., Stephen C. Porter, and Daniel B. Botkin. *The Blue Planet: An Introduction to Earth System Science.* 2nd. ed. New York: John Wiley and Sons, 1999.

THE STUDY OF EARTH

STUDYING EARTH

CONCEPT

The physical sciences include astronomy, physics, chemistry, and the earth sciences, but the last of these sciences is quite unlike the other three. Whereas the objects of study in physics and chemistry often seem abstract to the uninitiated and astronomy is concerned with faraway planets and other bodies, the earth sciences are devoted to things that are both concrete and immediate. The focus of study for the earth sciences is literally underneath our feet, a planet at once vast and tiny, a world that (as far as we know) stands alone in the universe as the sole supporter of intelligent life. The earth sciences also differ from other disciplines in that their boundaries are not always defined clearly. The study of Earth is a multifarious array of specialties that includes a range of geologic, hydrologic, and atmospheric sciences that overlap with the other physical sciences, biology, and even the social sciences.

HOW IT WORKS

Introduction to the Earth Sciences

At the simplest level, the earth sciences can be divided into three broad areas: the geologic, hydrologic, and atmospheric sciences. These specialties fit neatly with three of the "spheres," or subsystems within the larger Earth system: geosphere, hydrosphere, and atmosphere. (See Earth Systems for more about the spheres.) The fourth of these subsystems is the biosphere, and this illustrates the difficulty of stating exactly what is and what is not a part of the earth sciences.

Usually the earth sciences are considered part of the physical sciences, as opposed to the biological sciences, such as biology, botany, and zoology. Yet the earth sciences clearly overlap with biological sciences in a variety of areas, such as oceanography and various studies of complex biological environments. There is even a new (and as yet not fully formalized) discipline called geophysiology, built on the premise that Earth has characteristics of a living organism.

OVERLAP WITH OTHER PHYSICAL SCIENCES. The earth sciences also overlap with other physical sciences in several realms. There is geophysics, which addresses the planet's physical processes, including its magnetic and electric properties and the means by which energy is transmitted through its interior. There is also geochemistry, which is concerned with the chemical properties and processes of Earth. And there are numerous areas of confluence between the earth sciences and astronomy (among them, planetary geology), which fall under the heading of planetary science (sometimes called planetology or planetary studies).

These terms all refer to the same discipline, a branch of the earth sciences concerned with the study of other planetary bodies. This discipline, or set of disciplines, is concerned with the geologic, geophysical, and geochemical properties of other planets but also draws on aspects of astronomy, such as cosmology. Regardless of the name by which it is called, planetology is an example of the fact that the study of Earth is still very much an evolving set of disciplines. In many cases, the earth sciences are still in process of being defined.

Scientific Paradigms

This last point is an important one to consider because of what it implies about the nature of scientific study. In the past, scientists tended to think that they were in the business of discovering some sort of objective truth that was waiting for them to discover it; in reality, however, the quest of the scientific thinker is much less guided. The natural world does not in any way speak to the scientist, telling him or her how to categorize data. In fact, the divisions of scientific knowledge with which we are familiar have come about not because they necessarily reflect an underlying truth, but because they have proved useful in separating certain aspects of the physical world from certain others.

When science had its beginnings in ancient times, scientists were simply collecting observations (including a lot of incorrect ones) and sometimes forming theories of a sort, but they did not think in terms of developing models for viewing their objects of study. Today, however, scientific thinkers are acutely conscious of the model, or paradigm, that governs a particular discipline, school of thought, or theory.

A paradigm may be likened to a lens. The lens does not change the actual object that is viewed through it; it can alter only the way in which it is viewed. As thinkers within a particular discipline or theory begin to define the governing paradigm, they are much like an eye doctor testing various lenses on a patient. In such a situation, there is no one lens that is right for all circumstances. Rather, it is a question of finding the lens that best suits the patient's vision needs.

All sciences are gradually changing, evolving models that better suit the data under their consideration. Chemistry, for instance, was once primarily a matter simply of mixing chemicals and observing their external processes. In fact, the definition of chemistry has expanded greatly since about 1800, and today it is more like what people tend to think of as physics; that is, it is concerned with atomic and subatomic structures and types of behavior. The earth sciences are in an even more transitional state, and the problem of defining the disciplines it comprises is a still more fundamental one.

THE EVOLVING EARTH SCIENCES. In the discussion that follows, we outline the broad parameters of the earth sciences, considering basic areas of study and specialties within them. This does not represent a definitive organizational scheme, nor does this brief review refer to every possible area of study in the wide-ranging earth sciences. To do so would require an entire book; rather, the purpose is to consider the most significant disciplines and subdisciplines.

To appreciate the way that these disciplines fit within the larger perspective, however, it is necessary to examine a few historical details. Most of these details concern the early history of the earth sciences, since much of the more modern history (for example, the development of plate tectonics theory during the 1960s) is treated in the relevant essays within this book. The purpose of this brief historical review, instead, is to impart an understanding about how the study of Earth emerged as a real science, as opposed to a merely descriptive undertaking concerned with recording observations. Important themes are the development of the scientific method as well as the search for proper ways of classifying the various studies under the heading of what became known as the earth sciences.

REAL-LIFE APPLICATIONS

The Scientific Method

As discussed later, the scientific method emerged during the seventeenth century and has remained in use ever since. It is a way of looking at facts and data, and its application is what truly separates science from nonscience. Nonscientific "theories" postulate answers based not on evidence but on pure conjecture, a habit of mind that was widespread before the development of the scientific method and is still all too common. A contemporary example would be the claim that intelligent extraterrestrial life-forms built the great pyramids of Egypt.

The only real basis for this belief is the fact that the pyramids are extremely sophisticated architectural achievements for a civilization that had no metal tools, no wheel, and virtually no understanding of geometry, as was the case in Egypt in about 2500 B.C. But is a huge conceptual leap to go from the observation of these anomalies to the claim that visitors from outer space built the pyramids. The same is true of other strange artifacts from ancient or prehistoric times, such as the great structures of Stonehenge in England or the Nazca lines in South America. They are curious, puzzling, and intriguing, and it may be fun to speculate about engineers from another world—but such speculation is *not* science. (For more about the methodological distinctions between science and its opposite, see Earth, Science, and Nonscience.)

It is no mistake that here we are talking about the application of the scientific method to something outside the "hard sciences," the study of the pyramids being the province of social sciences, such as archaeology and history. In fact, the method has had just as much impact in those areas as it has in the physical and biological sciences. The scientific method (along with the closely related philosophical principles of basic logic, handed down from the ancient Greeks) can be applied in many aspects of daily life, enabling a person to make sense of a complex world. Many of the controversies of the modern world, including those involving race, sex, religion, and politics, could be treated more constructively if people approached the topics with a genuine interest in understanding the facts rather than simply finding confirmation for their emotionally based preconceived notions.

APPLYING THE SCIENTIFIC METHOD. Scientists rigorously applying the scientific method begin their quest for understanding by looking solely at the facts that can be garnered by observation. On the basis of these data, they form hypotheses, or unproven statements regarding observed phenomena. This is usually as far as many people go in their thinking, and it is not far enough: up to this point, for instance, the theory that claims that visitors from another planet built the pyramids is in accordance with the scientific method. But such fanciful notions, as well as kindred ideas, such as conspiracy theories, never go to the next step, which marks the dividing line between science and pure opinion.

Having formed a hypothesis, the scientist subjects it to testing, the most critical component of the scientific method. By contrast, advocates of nonscientific ideas (which, of course, usually pose as scientific ideas) typically focus on searching for evidence that will confirm the hypothesis. Anything that supports the hypothesis is reported; anything that does not is simply ignored. By contrast, a true scientist is constantly trying to *disprove* his or her own hypotheses.

Nicolaus Copernicus *(Library of Congress.)*

If a hypothesis withstands enough repeated testing, it acquires the status of a theory, or a more general statement about nature. If the idea embodied in a theory is shown to be the case in every situation for which it is tested, it then becomes a law. The process is a bit like that involved in making metal stronger: the more abuse it endures, assuming it is able to recover, the more impervious it will be to further abuse. But every type of metal has its limits, a threshold of compression or tension beyond which it cannot retain its original shape, and likewise it is possible that a scientific law can be overturned. For this reason, all laws are subject to continual testing, and if a test disproves a scientific law, this opens the way for the development of a new paradigm.

The Historical Roots of the Earth Sciences

The earth sciences are both old and new. On the one hand, they address matters of fundamental importance to human beings, and for this reason, the rudiments of the earth sciences probably made their appearance before any of the other fields of scientific study except perhaps astronomy. From prehistoric times, societies have been concerned with obtaining metals from the ground to make tools and weapons, finding water to support human life and crops, and discerning the future of weather patterns that could greatly affect conditions for human populations. Thus were born, respectively, the geologic, hydrologic, and atmospheric sciences.

Much of what took place in the earth sciences before about 1800, however, was a matter of superstition, legend, guesswork, and a smattering of real science. Much of it was dominated by religious belief, which relied on a strict interpretation of the Bible. Based on the amount of time that elapsed between Adam and Jesus, combined with the fact that Genesis 1 states that Adam was created at the end of the first week of Earth's existence, the Catholic Church maintained that the planet could not possibly be more than a few thousand years old. (For more on this subject, see Earth, Science, and Nonscience.)

THE ANCIENT EARTH SCIENCES. Despite the many impediments to scientific study in ancient times, a few thinkers contributed significantly to our knowledge. For instance, the Greek philosopher Aristotle (384–322 B.C.) discovered that Earth is a sphere by noting the rounded shadow on the Moon during a lunar eclipse. His pupil, Theophrastus (372?–287? B.C.), wrote a highly competent work, *Concerning Stones,* that remained a guide to mineralogy for two millennia. A few centuries later, the Greek mathematician Eratosthenes of Cyrene (*ca.* 276–*ca.* 194 B.C.) made an astoundingly accurate measurement of Earth's circumference.

Much of what passed for science, however, was little more than entertaining, anecdotal misinformation. Such is the case, for instance, in the *Historia Naturalis* (Natural history) of the Roman scholar Pliny the Elder (A.D. 23–79), a work that, despite its many flaws, remained widely respected through the Renaissance. As for Eratosthenes' measurement of Earth, the Alexandrian astronomer Ptolemy (*ca.* A.D. 100–170) rejected it in favor of a much smaller, much less correct figure. Thus, Ptolemy may deserve some of the credit for discovering the New World: if Christopher Columbus (1451–1506) had known just how far it was around Earth, he might not have been so confident about sailing off into the seas to the west of Europe.

Ptolemy's rejection of Eratosthenes's measurement was far from his only negative contribution to the history of science. Influenced by highly misguided concepts handed down from Aristotle himself (see Earth, Science, and Nonscience), he developed a complex cosmology that depicted Earth as the center of the universe. By this he meant that Earth was the center of the solar system, because up until a few centuries ago, astronomers believed that space consisted only of Earth, Sun, Moon, the five planets visible to the naked eye, and the "fixed stars" in the night sky.

DAWN OF THE SCIENTIFIC METHOD. What made Ptolemy's cosmology so complex, of course, was the fact that Earth is not anywhere near the center of the universe, and therefore his system required intricate mathematical acrobatics to remain workable. This posed little problem during the early Middle Ages, when learning in Europe all but ceased, and even in the much more scientifically progressive Muslim world of that time, Ptolemy's word remained virtual holy writ. By the late Middle Ages, however, as scientific learning returned to Europe, thinkers began to notice increasing difficulties in using his system.

The watershed event in what became known as the Scientific Revolution was the proof, by the Polish astronomer Nicolaus Copernicus (1473–1543), that Earth and the other planets of the solar system revolve around the Sun. By that point, the Catholic Church had given its official approval to Ptolemy's geocentric model, because it comported well with the idea that God had created humankind in his own image to fulfill a specific destiny. Therefore, Copernicus's challenge to established teachings proved highly controversial, and the Italian astronomer Galileo Galilei (1564–1642) would be forced to recant his support of it or face punishment by death.

Yet Galileo paved the way for the full acceptance of Copernicus's work and for the Scientific Revolution that followed in its wake. His theories and experiments concerning gravitational acceleration greatly influenced the English natural philosopher Isaac Newton (1642–1727), leading to the latter's epochal work on gravitation and motion. But at least as important as Galileo's work was his methodology: Galileo virtually introduced the scientific method, providing a set of principles for systematic study.

The Foundations of Modern Geology

The scientific method had an enormous impact on all the sciences. Unlike earlier "scientific" principles, which were built on the teachings of religious prophets or the uninformed conjecture of philosophers, this one was established on a foundation of observation, and it opened the way for unprecedented progress in the sciences.

Until late in the eighteenth century, however, the relatively young field of geology centered primarily on mere observation rather than the development of theories. Thus, the discipline was not all that different from what it had been in ancient times, or when the Anglo-Saxon historian known as the Venerable Bede (673–735) coined the term geology. The latter term, a combination of the Greek *geo* and *logia,* means "study of Earth," and was intended to distinguish such pursuits from theology, or the study of heavenly things.

HISTORICAL AND PHYSICAL GEOLOGY. In modern times, geology is defined as the study of the solid earth, in particular, its rocks, minerals, fossils, and land formations. As for Bede's putative opposition between geology and theology, it would become more pronounced in the period from about 1500 to 1800, as the findings of geologists began increasingly to contradict the teachings in the biblical book of Genesis. Among the first to consider the age of Earth in scientific terms, rather than by recourse to the Scriptures, was one of the world's greatest thinkers: the Italian scientist and artist Leonardo da Vinci (1452–1519), who speculated that fossils might have been made by the remains of long-dead animals.

Less famous was Leonardo's German contemporary Georgius Agricola (1494–1555), the "father of mineralogy," who wrote extensively on mining, metallurgy, and minerals. Together, these two men represent the two principal strains of geology: historical geology, or the study of Earth's history, and physical geology. The latter discipline, of which Agricola was a key representative, is concerned with the material components of Earth and with the forces that have shaped the planet. All the areas of geology discussed here fall under one of those two headings.

Most of the important developments in geology during the period from 1500 to 1800 fall under the heading of historical geology, beginning with a key observation on strata, or layers of rock, made by the Danish geologist Nicolaus Steno (1638–1687). As Steno correctly hypothesized, the lower a layer of rock lies, the earlier the historical period it represents. These observations, later developed into a theory by the German geologist Johann Gottlob Lehmann (1719–1767), had several implications.

First of all, the ideas of Steno and Lehmann provided geologists with a method for dating the age of rock formations not unlike the rings observed by dendrochronologists studying the biography of a tree. As a result of study based on these findings, scientists were confronted with the growing realization that Earth is much, much older than a strict interpretation of the Bible would suggest. This finding, in turn, led to the first theories concerning the shaping of Earth and thus to the foundation of geology as a modern scientific discipline.

THREE IMPORTANT SCHOOLS OF THOUGHT. In the wake of this breakthrough, at least three schools of thought developed. One of them, catastrophism, centered around the foregone conclusion that Earth had been created in six literal days or, at the very least, in an extremely short time, through a series of catastrophes. Opposed to this view was the Neptunist stratigraphy of the German geologist Abraham Gottlob Werner (1750–1817), who maintained that Earth had been shaped by a vast ocean (hence the name *Neptune*) that once covered its entire surface. Finally, there was the Plutonist school of the Scottish geologist James Hutton (1726–1797). Named after the Greek god of the underworld, this theory held that volcanoes and other disturbances beneath Earth's surface had been the principal forces in shaping the planet.

Hutton's theory would prevail, and today he is regarded as the father of modern geology. In *Theory of the Earth* (1795), he introduced one of the key concepts underlying the study of the planet's history, the principle of uniformitarianism—the idea that the forces at work on Earth today have always been in operation and are the same ones that shaped it. Nonetheless, Neptunism and even catastrophism had their merits. Although Werner and his followers were incorrect, Neptunism was the first well-developed theory concerning Earth's origins and helped pave the way for others.

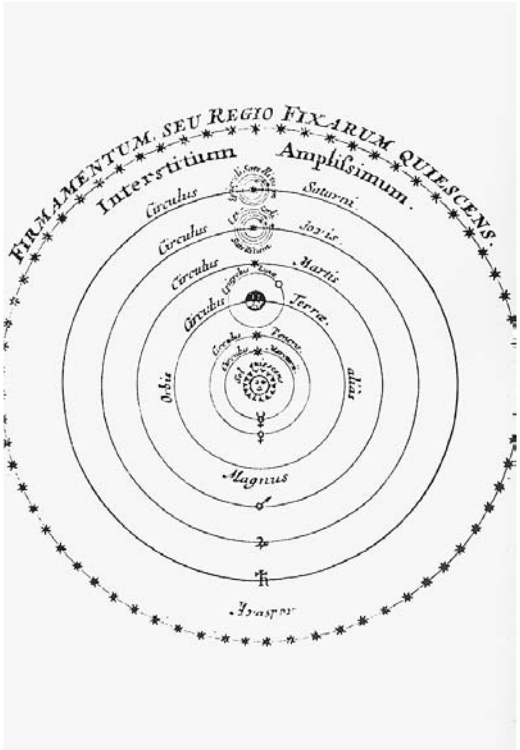


DIAGRAM SHOWING THE SOLAR SYSTEM AS COPERNICUS ENVISIONED IT, WITH THE SUN AT THE CENTER. THE COPERNICAN SYSTEM HERALDED THE START OF THE SCIENTIFIC REVOLUTION. *(© Dr. Jeremy Burgess/Photo Researchers. Reproduced by permission.)*

As for the advocates of catastrophism, they were correct inasmuch as they noted the role of sudden catastrophes in shaping the planet. These catastrophes (for instance, a comet about 66 million years ago, which may have destroyed the dinosaurs), however, can be explained within the framework of a very old Earth. Nor does the fact of the catastrophes themselves in any way suggest a planet that is only a few thousand years old.

LATER DEVELOPMENTS. As noted earlier, the later history of the earth sciences is discussed more properly within the context of specific subjects. Instead, our focus here is on the array of disciplines that proliferated alongside geology and on the need for a disciplinary paradigm larger than that of geology alone. By the mid–twentieth century, the range of disciplines involved in the study of Earth had become so complex and varied that it was a major achievement when the English geologist Arthur Holmes (1890–1965) developed a model that incorporated most of them. Holmes's system was not simply a "model" in the way that the term typically is used; Holmes also constructed a literal diagram that enabled students to visualize the relationships between subdisciplines.

Holmes's model was concerned primarily with the solid earth sciences, or the geologic sciences, meaning that it did not include the hydrologic sciences. Within its purview, however, it used a method of classification so broad (yet still targeted) that it has been adapted in recent years to include subdisciplines developed since Holmes's time. These changes serve to emphasize further the evolving nature of what came to be known as the earth sciences. The latter term came into use only during the 1960s and 1970s, when it became apparent that neither geology alone nor even a combination of geology, geophysics, and geochemistry could encompass all the areas of study devoted to Earth.

OVERVIEW OF THE EARTH SCIENCES

Throughout most of what remains of this essay, we very briefly sketch the outlines of the earth sciences. It should be reiterated that the organizational system used here is not necessarily definitive and is intended only to provide the reader with a general idea as to how the various earth sciences fit together.

GEOLOGY. At the core of the earth sciences, of course, is geology itself, which focuses on the study of the solid earth. As noted earlier, geology can be subdivided into historical and physical geology. The principle subdisciplines of historical geology are as follows.

- Stratigraphy: the study of rock layers, or strata, beneath Earth's surface
- Geochronology: the study of Earth's age and the dating of specific formations in terms of geologic time
- Sedimentology: the study and interpretation of sediments, including sedimentary processes and formations
- Paleontology: the study of fossilized plants and animals, or flora and fauna
- Paleoecology: the study of the relationship between prehistoric plants and animals and their environments.

Note that there are several other disciplines referred to by the prefix *paleo-* (or *palaeo-*), Greek for "very old." Two of the more well-known ones are paleobiology and paleobotany, but the subdisciplines can become very specialized, as evidenced by the existence of a field

KEY TERMS

BIOSPHERE: A combination of all living things on Earth—plants, mammals, birds, reptiles, amphibians, aquatic life, insects, viruses, single-cell organisms, and so on—as well as all formerly living things that have not yet decomposed. Typically, after decomposing, a formerly living organism becomes part of the geosphere.

COSMOLOGY: The study of the origin, structure, and evolution of the universe.

ECONOMIC GEOLOGY: The study of fuels, metals, and other materials from Earth that are of interest to industry or the economy in general.

GEOCENTRIC: Earth-centered.

GEOCHEMISTRY: A branch of the earth sciences, combining aspects of geology and chemistry, that is concerned with the chemical properties and processes of Earth.

GEOCHRONOLOGY: The study of Earth's age and the dating of specific formations in terms of geologic time.

GEOLOGY: The study of the solid earth, in particular, its rocks, minerals, fossils, and land formations.

GEOMORPHOLOGY: The study of landforms and of the forces and processes that have shaped them.

GEOPHYSICS: A branch of the earth sciences that combines aspects of geology and physics. Geophysics addresses the planet's physical processes as well as its magnetic and electric properties and the means by which energy is transmitted through its interior.

GEOSPHERE: The upper part of Earth's continental crust or that portion of the solid earth on which human beings live and which provides them with most of their food and natural resources.

HISTORICAL GEOLOGY: The study of Earth's physical history. Historical geology is one of two principal branches of geology, the other being physical geology.

HYDROSPHERE: The entirety of Earth's water, excluding water vapor in the atmosphere but including all oceans, lakes, streams, groundwater, snow, and ice.

HYPOTHESIS: An unproven statement regarding an observed phenomenon.

LAW: A scientific principle that is shown to always be the case and for which no exceptions are deemed possible.

MINERALOGY: The study of minerals (crystalline structures that make up rocks), which includes several smaller subdisciplines, such as crystallography.

PALEONTOLOGY: The study of fossilized plants and animals, or flora and fauna.

PHYSICAL GEOLOGY: The study of the material components of Earth and of the forces that have shaped the planet. Physical geology is one of two principal branches of geology, the other being historical geology.

PHYSICAL SCIENCES: Astronomy, physics, chemistry, and the earth sciences.

PLANETARY SCIENCE: The branch of the earth sciences, sometimes called planetology or planetary studies, that focuses on the study of other planetary bodies. This discipline, or set of disciplines, is concerned with the geologic, geophysical, and geochemical properties of other

KEY TERMS CONTINUED

planets but also draws on aspects of astronomy, such as cosmology.

SCIENTIFIC METHOD: A set of principles and procedures for systematic study that includes observation; the formation of hypotheses, theories, and laws; and continual testing and reexamination.

SCIENTIFIC REVOLUTION: A period of accelerated scientific discovery that completely reshaped the world. Usually dated from about 1550 to 1700, the Scientific Revolution saw the origination of the scientific method and the introduction of ideas such as the heliocentric (Sun-centered) universe and gravity.

SEDIMENTOLOGY: The study and interpretation of sediments, including sedimentary processes and formations.

STRATIGRAPHY: The study of rock layers, or strata, beneath Earth's surface.

STRUCTURAL GEOLOGY: The study of rock structures, shapes, and positions in Earth's interior.

THEORY: A general statement derived from a hypothesis that has withstood sufficient testing.

known as paleobiogeography, or the study of fossils' geographic distribution. The principle subdisciplines of physical geology are:

- Geomorphology: the study of landforms and of the forces and processes that have shaped them
- Structural geology: the study of rock structures, shapes, and positions in Earth's interior
- Mineralogy: the study of minerals (crystalline structures that make up rocks), which includes several smaller subdisciplines, such as crystallography
- Petrology: the study of rocks, which is divided into several smaller subdisciplines, most notably igneous, metamorphic, and sedimentary petrology
- Economic geology: the study of fuels, metals, and other materials from Earth that are of interest to industry or the economy in general
- Environmental geology: the study of the geologic impact of both natural and human activity on the environment.

It should be noted that there is some overlap between historical and physical geology. For instance, sedimentology often is placed under the heading of physical geology, while some sources include a third category of subdisciplines that overlap both historical and physical geology.

OTHER GEOLOGIC SCIENCES. Geology occupies a central place among the geologic sciences or geosciences, but also important are those disciplines and subdisciplines formed, as Holmes pointed out, at the intersections between geology and astronomy, physics, and chemistry, respectively. (Some sources, on the other hand, consider these disciplines to be a part of geology itself. In the present context, the term geologic sciences is used to encompass not only geology but also these related areas of study.)

Planetary science applies the earth sciences paradigm to other planets. Among its important subdisciplines is astrogeology or planetary geology, or the study of the rock record on the Moon, the planets, and other bodies. Also significant is cosmology, the study of the origin, structure, and evolution of the universe, which often is treated as part of astronomy.

Geophysics, or an application of physics to the study of Earth, occupies a position of prominence within the earth sciences. Among the areas it addresses are the production, expenditure, and transmission of energy within Earth as well as the planet's magnetic, electric, and gravitational properties. Geophysics encompasses such areas as geodesy, the science of measuring Earth's shape and gravitational field. Seismology, or the study of the waves produced by earthquakes and volcanoes, is another important part of geophysics. (On the other hand, volcanology, or the study of volcanoes themselves, would fall more properly under physical geology.)

Geochemistry, which is concerned with the chemical properties and processes of Earth, covers a wide array of natural phenomena—from radioactive isotopes in the ground to life-forms in the biosphere. Under the heading of geochemistry fall several biogeochemical processes, such as the carbon cycle, whose study brings together aspects of the physical sciences geology and chemistry as well as various life sciences.

OTHER EARTH SCIENCES. The hydrological sciences are concerned with the hydrosphere and its principal component, water. These disciplines include hydrology, the study of the water cycle; glaciology, the study of ice in general and glaciers in particular; and oceanography. Clearly, oceanography overlaps with the life sciences; likewise, hydrogeology (the study of groundwater), as its name implies, overlaps with geology.

The atmospheric sciences, obviously, are devoted to the atmosphere. Most notable among these sciences is meteorology, the study of weather patterns, and climatology, the study of temperature and climate. (Paleoclimatology is an important subdiscipline of historical geology.) The atmospheric sciences also are concerned with phenomena ranging from pollution to the optical effects created by the interaction of the Sun's rays with the atmosphere.

Finally, there are miscellaneous areas of study that either are interdisciplinary or cross boundaries between the earth sciences and the social sciences. In the former category, for instance, would be environmental studies that involve aspects of the biosphere, atmosphere, geosphere, and hydrosphere. Examples of the second category are paleoarchaeology, the study of the earliest humans and humanoid forms, and, of course, geography. Also included in this group are such intriguing areas as urban geology, a branch of environmental geology concerned with human settlements.

WHERE TO LEARN MORE

Allaby, Ailsa, and Michael Allaby. *A Dictionary of Earth Sciences.* 2d ed. New York: Oxford University Press, 1999.

Athro: Your Source for High School and College Level Biology, Earth Science, and Geology on the Web (Web site). <http://www.athro.com/>.

Cox, Reg, and Neil Morris. *The Natural World.* Philadelphia: Chelsea House Publishers, 2000.

Dasch, E. Julius. *Earth Sciences for Students.* New York: Macmillan Reference USA, 1999.

Geology Entrance (Web site). <http://www.ucmp.berkeley.edu/exhibit/geology.html>.

Hancock, Paul L., and Brian J. Skinner. *The Oxford Companion to the Earth.* New York: Oxford University Press, 2000.

Illustrated Glossary of Geologic Terms (Web site). <http://www.geology.iastate.edu/new_100/glossary.html>.

Jobs in Earth Sciences (Web site). <http://geology.com/jobs.htm>.

Knapp, Brian J., David Woodroffe, and Julian Baker. *Earth Science: Discovering the Secrets of the Earth.* Danbury, CT: Grolier Educational, 2000.

Skinner, Brian J., Stephen C. Porter, and Daniel B. Botkin. *The Blue Planet: An Introduction to Earth System Science.* 2d ed. New York: John Wiley and Sons, 1999.

Stace, Alexa. *Atlas of Earth.* Milwaukee, WI: Gareth Stevens Publishing, 2000.

MEASURING AND MAPPING EARTH

CONCEPT

Today a sharp distinction exists between the earth sciences and geography, but this has not always been the case. In ancient times, when scientists lacked the theoretical or technological means to study Earth's interior, the two disciplines were linked much more closely. Even in the centuries since these disciplines parted ways, the earth sciences have continued to benefit from a foundation established in part by early geographers, whose work informed the geophysical subdiscipline of geodesy. Like geographers, earth scientists are interested in measuring and mapping Earth, though their interests are quite different. Among the areas of concern to earth scientists are the location of underground resources and the obtaining of data on the planet's gravitational and magnetic fields. In these and other pursuits, earth scientists use a number of techniques and technologies, ranging from the ancient discipline of surveying to the most modern forms of satellite-based remote sensing.

HOW IT WORKS

GEOGRAPHY AND GEOLOGY

Whereas geology is the study of the solid earth, including its history, structure, and composition, geography is the study of Earth's surface. Geologists are concerned with the grand sweep of the planet's history over more than four billion years, whereas the work of geographers addresses the here and now—or at most, in the case of historical geography, a span of just a few thousand years.

Today the distinctions between these two disciplines are sharp, so much so that most books on the earth sciences barely even mention geography. In a modern university, chances are that the geography and geology/earth sciences departments will not even be located in the same building. Geography, after all, usually is classified among social sciences such as anthropology or archaeology, whereas geology is a "hard science," along with physics or biology.

It is true that the divisions between geography and geology are clear, symbolized by the features that appear or do not appear on the maps used by either discipline. Geography is somewhat concerned with natural features, but its interests include man-made boundaries, points, and such formations as population centers, roads, and so on. Certainly, an atlas may include physical maps, which are dominated by natural features and contain little or no evidence of man-made demarcations or points of interest. Nonetheless, the purpose of a geographical atlas is to identify locations of interest to humans, among them, cities, roads from one place to another, and borders that must be crossed.

By contrast, geologic maps contain detailed information about rock formations and other natural features, with virtually nothing to indicate the presence of humans except as it relates to natural features under study. An exception might be a map designed to be used by paleoarchaeologists, who study the earliest humans and humanoid forms. Their discipline, which combines aspects of the earth sciences and archaeology, is concerned with human settlements, but mostly only prehistoric human settlements.

Despite the depth and breadth of distinctions between them, it is significant that studies in both geography and geology make use of maps. Mapmaking, or cartography, is considered a subdiscipline of geography, yet it is actually an interdisciplinary pursuit (much like many of the earth sciences—see Studying Earth) and combines aspects of science, mathematics, technology, and even art. Although their interests are in most cases quite different from those of geographers, geologists rely heavily on the work of cartographers.

EARLY GEOGRAPHIC STUDIES

The history of the sciences has been characterized by the continual specialization and separation of disciplines. Thus, it should not be surprising to discover that to the ancients, the lines were blurred between geography, mathematics, astronomy, and what people today would call earth sciences. Most of the early advances in the study of Earth involved all of those disciplines, an example being the remarkable estimate of Earth's size made by Eratosthenes of Cyrene (*ca.* 276–*ca.* 194 B.C.).

A mathematician and librarian at Alexandria, Egypt, Eratosthenes discovered that at Syene, several hundred miles south along the Nile near what is now Aswan, the Sun shone directly into a deep well and upright pillars cast no shadow at noon on the summer solstice (June 21). By using the difference in angles between the Sun's rays in both locations as well as the distance between the two towns, he calculated Earth's circumference at about 24,662 mi. (39,459 km). This figure is amazingly close to the one used today: 24,901.55 mi. (39,842.48 km) at the equator. Eratosthenes published his results in a book whose Greek name, *Geographica* (Geography), means "writing about Earth." This was the first known use of the term.

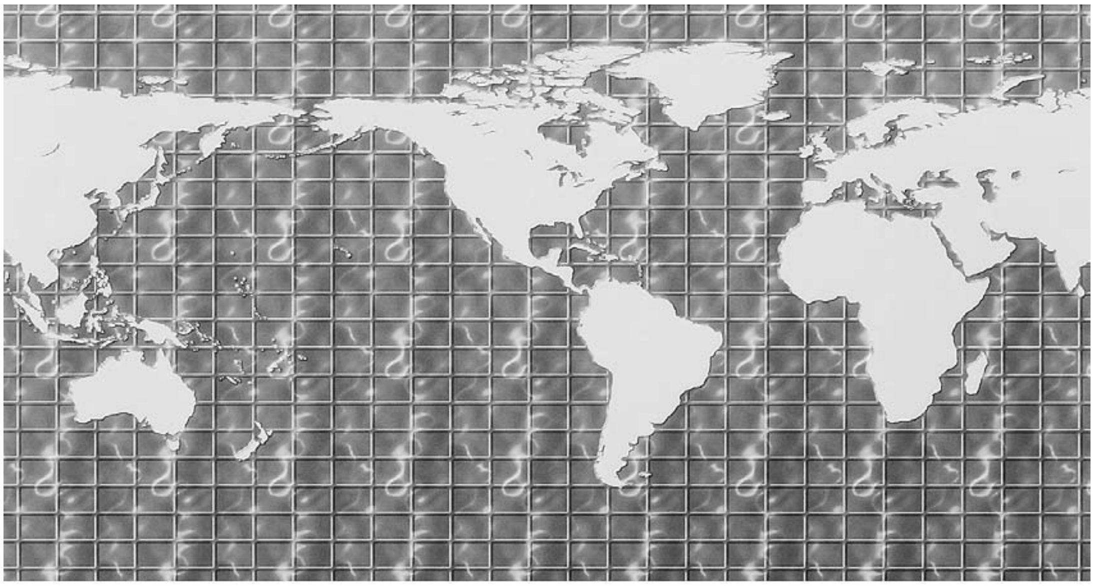
PTOLEMY'S GEOGRAPHY. Unfortunately, the Alexandrian astronomer Ptolemy (*ca.* A.D. 100–170), one of the most influential figures of the ancient scientific world, rejected Eratosthenes' calculations and performed his own, based on faulty information. The result was a wildly inaccurate estimate of 16,000 mi. (25,600 km). More than thirteen centuries later, Christopher Columbus (1451–1506) relied on Ptolemy's figures rather than those of Eratosthenes, whose work was probably unknown to him. Thinking that the circumference of Earth was two-thirds what it actually is, Columbus set sail westward from Spain—something he might not have done if he had known about the "extra" 8,000 mi. (12,874 km) that lay to the west of Europe.

Nonetheless, in his *Hyphegesis geographike* (Guide to geography), Ptolemy did make useful contributions to geographical study. He helped popularize the use of latitude and longitude lines, first conceived by the Greek astronomer and mathematician Hipparchus (*fl.* 146–127 B.C.), and rejected the widespread belief that a vast ocean, known to the ancients as "the Ocean Sea," surrounded the entire world. He also presented a set of workable mathematical principles for representing the spherical surface of Earth on a flat page, always a problem for cartographers.

In addition, Ptolemy established the practice of orienting maps with north at the top of the page. Today this is taken for granted, but in his time cartographers depicted the direction of the rising Sun, east, at the top of their maps. Ptolemy used a northward orientation because the Mediterranean region that he knew extended twice as far east to west as it did north to south. To represent the area on a scroll, the form in which books appeared during his time, it was easier to make maps with north at the top.

OTHER ANCIENT GEOGRAPHERS. Most other ancient geographers of note were primarily historians rather than scientists, preeminent examples being the Greek Herodotus (*ca.* 484–*ca.* 430–420 B.C.), the father of history, and Strabo (*ca.* 64 B.C.–*ca.* A.D. 24), whose work provides some of the earliest Western descriptions of India and Arabia. An exception was Pomponius Mela (*fl. ca.* A.D. 44), whose work would fit into the subdiscipline known today as physical geography, concerned with the exterior physical features and changes of Earth.

In his three-volume *De situ orbis,* (A description of the world), Mela introduced a system of five temperature zones: northern frigid, northern temperate, torrid (very hot), southern temperate, and southern frigid. Unlike many scientific works of antiquity, Mela's geography has remained influential well into modern times, and his idea of the five temperature zones remains in use. Mela and Ptolemy, who followed him by about a century, were among the last geographers



Mercator cylindrical projection of the continents of Earth, showing the characteristic exaggeration in scale of land masses near the poles. *(© R. Winter/Photo Researchers. Reproduced by permission.)*

of note in the West for more than a thousand years.

THE SEPARATION OF GEOGRAPHY AND EARTH SCIENCES. During the first half of the Middle Ages significant work in cartography took place in the Muslim world and in the Far East, but not in Europe. Only in the course of the Crusades (1095–1291) did Europeans become interested in exploration again, and this interest grew as the Mongol invasions of the thirteenth century opened trade routes from Europe to China for the first time in a thousand years. The crusading and exploring spirits met in Henry the Navigator (1394–1460), the prince who, while he never actually took part in any voyages himself, quite literally launched the Age of Exploration from his navigation school at Sagres, Portugal.

In the two centuries that followed, European mariners and conquerors explored and mapped the continents of the world. Along the way, these nonscientists sometimes added knowledge to what would now be considered the earth sciences, as when the Italian explorer Christopher Columbus (1451–1506) became the first to notice magnetic declination. (See Geomagnetism for more on this subject.) Meanwhile, Columbus's contemporary, the Italian artist and scientist Leonardo da Vinci (1452–1519), as well as Georgius Agricola (1494–1555) of Germany, known as the father of mineralogy, conducted some of the first modern studies in the earth sciences. (See Studying Earth for more about Leonardo's and Agricola's contributions.)

In the sixteenth century, the Flemish cartographer Gerhardus Mercator (1512–1594) greatly advanced the science of mapmaking with his development of the Mercator projection. The latter method, still used in many maps today, provided an effective means of rendering the spherical surface of Earth on a two-dimensional map. By that time, cartography had emerged as a vital subdiscipline, and over the ensuing two centuries the separation between geography and the earth sciences became more and more distinct.

The full separation of disciplines took place in the eighteenth century, when the German geographer Anton Friedrich Büsching (1724–1793) pioneered modern scientific geography. Beginning in 1754, Büsching published the 11-volume *Neue Erdbeschreibung* (New description of the earth), which established a foundation for the study of geography in statistics rather than descriptive writing. During the same century geology was in the middle of its own paradigm shift. Until the time of the Scottish geologist James Hutton (1726–1797), a near exact contemporary of Büsching, geologists had been concerned primarily with explaining how the world began—a topic that almost inevitably led to conflict over religious questions (see Earth, Science, and Nonscience). Hutton, known as the father of geology, was the first to transcend this issue and instead offer a theory about the workings of Earth's internal mechanisms.

SURVEYING

The era of Büsching and Hutton coincided with that of the English astronomer Charles Mason (1730–1787) and the English surveyor Jeremiah Dixon (*d.* 1777), who in 1763 began their famous survey of the boundary between Pennsylvania and Maryland. It took them five years to survey the 233-mi. (373-km) Mason–Dixon Line, which eventually became known as the border between the free and slave states before the Civil War. Surveying, a profession practiced by such great Americans as the first president, George Washington (1732–1799), and the mathematician and astronomer Benjamin Banneker (1731–1806), eventually became associated with the United States, but it originated in Egypt as early as 2700 B.C.

Surveying is a realm of applied mathematics devoted to measuring and mapping areas of land. Though it is obviously of value to the earth sciences and geography, it has a great deal of importance economically and politically as well. For this reason, the Romans, who were not nearly as inclined toward theoretical study as the Greeks, became preeminent surveyors whose land parcels still can be seen from the air over parts of western Europe. Owing to its great practical importance, surveying—unlike virtually all other forms of learning—continued to thrive in Europe during the early Middle Ages. Nonetheless, surveying improved in the Renaissance and thereafter, as the result of the introduction of new tools and mathematical techniques.

Among these tools were the theodolite, used for measuring horizontal and vertical angles, and the transit, a type of theodolite that employs a hanging plumb bob to determine a level sight line. Mathematical techniques included triangulation, whereby the third side of a triangle can be determined from measurements of the other two sides and angles. The German mathematician Karl Friedrich Gauss (1777–1855), regarded as the father of geodesy, introduced the heliotrope, a mechanism that aids in triangulation. Other tools included the compass, level, and measuring tapes; modern surveying benefits from remote sensing. (Both remote sensing and geodesy are discussed later in this essay.)

REAL-LIFE APPLICATIONS

GEOLOGIC MAPS AND SURVEYS

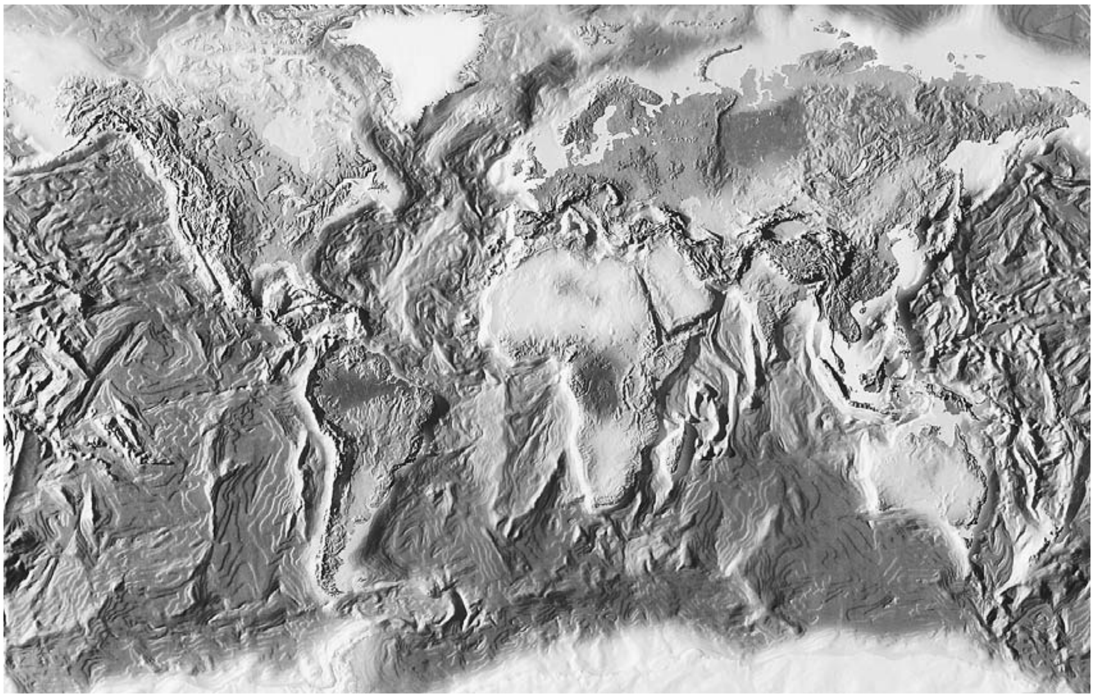
A geologic map shows the rocks beneath Earth's surface, including their distribution according to type as well as their ages, relationships, and structural features. The first geologic map, made in 1743, depicted subterranean East Kent, England. Its creator, the English physician and geologist Christopher Packe (1686–1749), introduced the technique of hachuring, that is, representing relief (elevation) on a map by shading in short lines in the direction of the slopes.

Three years later, the French geologist Jean-Etienne Guettard (1715–1786) made the first geologic map that crossed national lines, thus illustrating the distinction between geology and geography. As Guettard discovered, the geologic features of the French and English coasts along the English Channel are identical, indicating that the areas are connected. At a time when France and England were still bitter political and military rivals, Guettard's work showed that they literally shared some of the same land.

Geologic mapmaking received a further boost in 1815, when the English geologist William Smith (1769–1839) produced what has been called the first geologic map based on scientific principles. Entitled *A Delineation of the Strata of England and Wales with Part of Scotland,* the map used different colors to indicate layers of sedimentation. Roughly 6 ft. by 9 ft. (1.8 × 2.7 m), the map linked paleontology with stratigraphy (the study of fossils and the study of rock layers, respectively) and proved to be a milestone in geologic cartography.

MODERN GEOLOGIC MAPMAKING. Geologic mapmaking changed considerably in the period after World War II. Before that time, geologists did most of their work with the use of topographical maps, or maps that showed only surface features. Following the war, however, aerial photography became much more common, giving rise to the technique of photogeology, the use of aerial photographic data to make determinations regarding the geologic characteristics of an area.

Petroleum companies, which often have taken the lead in developing advanced methods of geologic study, introduced the practice of creating three-dimensional images from the air.



Relief-shaded map of Earth, showing the continents in raised elevation. *(© M. Agliolo/Photo Researchers. Reproduced by permission.)*

They did this by taking pairs of photographs which, when viewed through a stereoscope, provided images that could be studied in great detail for information about all manner of geologic features. Height proved a great advantage, revealing features that would not have been as clear to a geologist working on the ground.

Of course, a great deal of work on the ground was still necessary for confirming the data revealed by aerial surveillance and for other purposes, such as measurement and sample collection. Nonetheless, aerial photography provided an enormous boost to geologic studies, as did the use of satellite imaging from the 1970s onward. Also helpful were such new devices as the handheld magnetometer, which made it relatively easy to separate rocks containing magnetite from other, nonmagnetic samples.

WHY MAKE GEOLOGIC MAPS? One might wonder why any of this is important, aside from a purely academic interest in the structure of rocks under the ground. In fact, the need for precise geologic data goes far beyond "purely academic interests," as the reference to oil companies suggests. Geologic mapmaking is critical to the location of oil as well as minerals and other valuable natural resources—including the most useful one of all, water.

Likewise it is necessary to have accurate geologic information before undertaking a large engineering project, such as the building of a road or bridge. In such situations, geologic studies can quite literally be a matter of life and death, and eventually such studies may save more lives by aiding in the prediction of earthquakes or volcanoes. Geologic data is also a critical part of studies directed toward environmental protection, both for areas designed to remain natural habitats and for those designated for development. And, finally, there are studies whose purpose is purely, or mostly, academic but that reveal a great deal of useful information about the history of the planet, the forces that shaped it, and perhaps even future events.

GEOLOGIC SURVEYS. Geologic mapmaking is so vital, in fact, that national governments have undertaken large-scale and ongoing geologic studies since 1835. That was the year that Great Britain became the first country to establish a geologic survey, with the aim of preparing a geologic map of the entire British Isles. The project began with the mapping of Cornwall and southern Wales and has continued ever since, with the addition of new details as they have become available.

In the ensuing years several other countries established their own geologic surveys, including the United States in 1889. (The Web site of the U.S. Geological Survey is listed in the bibliography.) Other important national geologic surveys include those of France, Canada, China, and Russia. The last of these surveys is of particular interest, dating as it does from the "Stone Department," a mineralogical survey established in 1584. Today even much smaller nations, such as Uruguay, Slovakia, and Namibia, have their own national geologic surveys.

Over the years, techniques of information gathering have evolved, particularly with the development of satellite remote-sensing technology. So, too, have the areas under the purview of various national geologic surveys, which since the 1950s have undertaken the mapping of the continental shelves adjacent to their own shorelines. (In addition, the nations claiming territories in Antarctica, including Britain and the United States, have mapped the geologic features of that continent extensively, though mining or other economic development there is forbidden.) The U.S. Geological Survey has also seen its scope extended to include studies on such issues as radioactive waste disposal and prediction of natural hazards, among them, earthquakes in urban areas.

GEOPHYSICAL MEASUREMENTS

Geophysics is a branch of the earth sciences that combines aspects of geology and physics. Among the areas it addresses are Earth's physical processes as well as its gravitational, magnetic, and electric properties and the means by which energy is transmitted through its interior. Areas of geophysics with a particular focus on measurement and mapping include the study of geomagnetism, or Earth's magnetic field, and geodesy, which is devoted to the measurement of Earth's shape and gravitational field.

The measurement of gravitational fields involves the use of either weights dropped in a vacuum or mechanical force-balance instruments. The first of these techniques is much older than the other and provides an absolute measure of the gravitational field in a given area. As for force-balance instruments, they are similar in principle to scales and furnish a relative measure of the gravitational field. To compare the gravitational field at different positions, however, it is necessary to establish a frame of reference. This is known as the *geoid,* a surface of uniform gravitational potential covering the entire earth at a height equal to sea level. (See Gravity and Geodesy for more on these topics.)

GEODETIC MEASUREMENTS OF EARTH'S SURFACE. As noted, geodesy is concerned not only with Earth's gravitational field but also with its shape, and earth scientists working on this aspect of the subdiscipline employ many of the techniques and equipment described earlier with regard to geography and surveying. Eratosthenes's measurement of Earth's size is thought to be the first geodetic measurement, and in performing geodetic measurements today, earth scientists often employ concepts familiar to surveyors.

Among these concepts is triangulation, which was developed in the sixteenth century by the Dutch mathematician Gemma Frisius (1508–1555). Triangulation remained an important method of geodetic measurement until the development of satellite geodesy made possible simpler and more accurate measurements through remote sensing. Even today triangulation is still used by geologists without access to satellite data. In performing measurements using triangulation, geologists employ the theodolite, and typically at least one triangulation point is highly visible—for instance, the top of a mountain.

Until the 1950s scientists used a measuring tape of a material called Invar, a nickel-iron alloy noted for its tendency not to expand or contract with changes in temperature. From that time, however, electronic distance measurement (EDM) systems, which employ microwaves or visible light, came into use. EDM helped overcome some of the possibilities for error inherent in using any kind of tape, for example, the likelihood that it would sag and thus render incorrect measurements. Furthermore, EDM tended to reduce errors caused by atmospheric refraction.

With the advent of the United States program in the 1960s, increasingly more sophisticated forms of geodesic remote-sensing technology came into use. Among these techniques is satellite laser ranging, which relies on measurements of the amount of time required for a laser pulse to travel from a ground station to a satellite and back. Before the development of the global posi-

KEY TERMS

CARTOGRAPHY: The creation, production, and study of maps. Cartography is a subdiscipline of geography and involves not only science but also mathematics, technology, and even art.

DOPPLER EFFECT: The change in the observed frequency of a wave when the source of the wave is moving with respect to the observer.

FIELD: A region of space in which it is possible to define the physical properties of each point in the region at any given moment in time.

GEODESY: An area of geophysics devoted to the measurement of Earth's shape and gravitational field.

GEOGRAPHY: A social science concerned with the description of physical, biological, and cultural aspects of Earth's surface and with the distribution and interaction of these features. Compare with geology.

GEOID: A surface of uniform gravitational potential covering the entire earth at a height equal to sea level.

GEOLOGIC MAP: A map showing the rocks beneath Earth's surface, including their distribution according to type as well as their ages, relationships, and structural features.

GEOLOGY: The study of the solid earth, in particular, its rocks, minerals, fossils, and land formations.

GEOPHYSICS: A branch of the earth sciences that combines aspects of geology and physics. Geophysics addresses the planet's physical processes as well as its gravitational, magnetic, and electric properties and the means by which energy is transmitted through its interior.

GEOMAGNETISM: A term referring to the magnetic properties of Earth as a whole, rather than those possessed by a single object or place on Earth.

tioning system (GPS), discussed later, the use of satellite systems necessitated tracking through the Doppler effect, or the change in the observed frequency of a wave when the source of the wave is moving with respect to the observer. Thanks to GPS, put into operation by the U.S. Department of Defense, satellite tracking is much simpler and more accurate today.

GEOMAGNETIC MEASUREMENTS. Ever since the ancient Chinese discovered that pieces of lodestone (magnetite) tend to point north, mariners have used the compass for navigation. The compass was augmented by other navigational devices until it was supplanted by the gyroscope in modern times and still later by more sophisticated devices and methods, such as GPS. Yet a compass still works fine for many a hiker, and its use serves to emphasize the importance of Earth's magnetic field.

Earth has an overall geomagnetic field, and specific areas on the planet have their own local magnetic fields. Thanks in large part to the contributions of Gauss, who developed a standardized local magnetic coordinate system in the early nineteenth century, it became possible to perform reasonably accurate measurements of local magnetic data while correcting for the influence of Earth's geomagnetic field. Indeed, one of the challenges in measuring magnetic fields is the fact that the Earth system possesses magnetic force from so many sources: the molten core, from whence originates the preponderance of Earth's magnetic field; external fields, such as the magnetosphere and ionosphere; local materials, such as magnetite, hematite, or pyrrhotite;

KEY TERMS CONTINUED

HACHURING: A method of representing relief (elevation) on a map by shading in short lines in the direction of the slopes.

MAGNETOSPHERE: An area surrounding Earth, reaching far beyond the atmosphere, in which ionized particles (i.e., ones that have lost or gained electrons so as to acquire a net electric charge) are affected by Earth's magnetic field.

PALEOMAGNETISM: An area of historical geology devoted to studying the direction and intensity of magnetic fields in the past, as discerned from the residual magnetization of rocks.

PALEONTOLOGY: The study of fossilized plants and animals, or flora and fauna.

PHOTOGEOLOGY: The use of aerial photographic data to make determinations regarding the geologic characteristics of an area.

PHYSICAL GEOGRAPHY: A subdiscipline of geography concerned with the exterior physical features and changes of Earth.

POTENTIAL: Position in a field, such as a gravitational force field.

REFRACTION: The bending of light as it passes at an angle from one transparent material into a second transparent material.

REMOTE SENSING: The gathering of data without actual contact with the materials or objects being studied.

STRATIGRAPHY: The study of rock layers, or strata, beneath Earth's surface.

SURVEYING: An area of applied mathematics devoted to measuring and mapping areas of land.

TRIANGULATION: A technique in surveying whereby the third side of a triangle can be determined from measurements of the other two sides and angles.

and even man-made sources of magnetic or electric force.

After making calculations that correct for interfering sources of magnetism, geophysicists study the remaining magnetic anomalies, which can impart extremely valuable information. Classic examples include the discovery that Earth's magnetic polarity has reversed many times, a finding that led to the development of the geophysical subdiscipline known as paleomagnetism. Paleomagnetic studies, in turn, served as a highly significant confirmation of plate tectonics, which originated in the middle of the twentieth century and remains the dominant theory regarding geologic processes. (See Plate Tectonics. For more about geomagnetism, including some of the topics mentioned here, see Geomagnetism.)

Knowing One's Location

In these and other types of studies that involve mapping and measurement, it is important for scientists conducting surveys to be aware of the frame of reference from which they are operating—that is, the perspective from which they view data. Simply put, one must know first where one is before one can measure and map geophysical or other data for surrounding areas. This requires knowledge of latitude and longitude, or east–west and north–south positions, respectively.

From earliest times, mariners and scientists have been able to ascertain latitude with relative ease, simply by observing the angle of the Sun and other stars. Determination of longitude, however, proved much more difficult, because it required highly accurate timepieces. Only in the late eighteenth century, with the breakthroughs achieved by the British horologist John Harrison (1693–1776) did such calculations become possible. As a result, many a ship's crew was saved from the misfortunes that could result from inaccurate estimates of location.

GLOBAL POSITIONING SYSTEM. By the latter part of the twentieth century, navigational technology had become vastly more sophisticated than it was in Harrison's day. From the 1957 launch of the Soviet satellite *Sputnik 1,* the skies over Earth became increasingly populated with satellites, such that within half a century dozens of countries had payloads in space. Aside from governments and scientific research establishments, even cable television companies used satellites to beam programming to homes all over the industrialized world, a fact that in itself says much about the spread of satellite technology.

Among the most impressive uses of satellites is the GPS, developed by the U.S. Department of Defense to assist in surveillance. GPS consists of 24 satellites orbiting at an altitude of 12,500 mi. (20,000 km). They move in orbital paths such that an earthbound receiver can obtain signals from four or more satellites at any given moment. On board are atomic clocks, which provide exact time data with each signal and eliminate the necessity of the receiver's having such an accurate clock. By receiving data from these satellites, persons on the ground can compute their own positions in terms of latitude and longitude as well as altitude. Not all receivers have access to the most accurate data possible: in line with the strategic mission for which it initiated GPS in the first place, the Defense Department ensures that only authorized personnel receive the most precise information.

Thus, GPS has built-in errors so that civilian users can calculate locations with an accuracy of "only" 328-492 ft. (100–150 m). This, of course, is amazingly accurate, but not as accurate as the data available to those authorized to receive normally encrypted information on the P (Precise) code. The latter provides an accuracy of 3.28-16.4 ft. (1–5 m) instantaneously, and more detailed measurements based on GPS data can be used to achieve accuracy of up to 0.2 in. (5 mm).

Remote Sensing

Many of the methods used by geologists and geophysicists to map and measure Earth make use of remote sensing, the gathering of data without actual contact with the materials or objects being studied. Without remote sensing, it would be impossible to discuss many physical phenomena intelligently, because it is unlikely that any technology ever will make it possible to explore many areas underneath the planet's surface directly.

An example of remote sensing is photogeology, described briefly earlier; so, too, is satellite imaging for data collection. The earth scientist of the twenty-first century likewise has other highly sophisticated forms of technology, such as radar systems or infrared imaging, at his or her disposal. As with many other aspects of geologic mapping and measurement, this one has value far beyond the classroom: remote-sensing studies make it possible, for instance, to observe the environmental impact of deforestation in large geographical areas. (For much more about this subject, see Remote Sensing.)

WHERE TO LEARN MORE

Brooks, Susan. *The Geography of the Earth.* New York: Oxford University Press, 1996.

Erickson, John. *Exploring Earth from Space.* Blue Ridge Summit, PA: TAB Books, 1989.

Geologic Maps (Web site). <http://http://www.aqd.nps.gov/grd/usgsnps/gmap/gmap1.html>.

Hancock, Paul L., and Brian J. Skinner. *The Oxford Companion to the Earth.* New York: Oxford University Press, 2000.

Internet Resources for Geography and Geology (Web site). <http://www.uwsp.edu/geo/internet/geog_geol_resources.html>.

Moffitt, Francis H., and Harry Bouchard. *Surveying.* 8th ed. New York: Harper and Row, 1987.

Remote Sensing Data and Information (Web site). <http://rsd.gsfc.nasa.gov/rsd/RemoteSensing.html>.

U.S. Geological Survey (Web site). <http://www.usgs.gov/>.

Virtual Museum of Surveying (Web site). <http://www.surveyhistory.org/index.htm>.

Wilford, John Noble. *The Mapmakers.* New York: Knopf, 2000.

REMOTE SENSING

CONCEPT

Scientists of many disciplines are accustomed to studying data that cannot be observed through direct contact. Physicists and chemists, for instance, know a great deal about the structure of the atom, even though even the most high-powered microscope cannot make an atom visible to the human eye. The objects of study for earth scientists are often similarly remote, though not necessarily because they are small. In some cases, the problem is quite the opposite: an area selected for study is too large to provide understanding to geologists working only on the ground. Other areas are simply inaccessible to human beings or even their equipment. This has necessitated the development of remote sensing equipment and techniques, primarily involving views from the air or from space and utilizing electromagnetic radiation across a wide spectrum.

HOW IT WORKS

An Introduction to Remote Sensing

The work of geologists would be much easier if Earth were transparent and they could simply look down into the ground as they would into the sky. But the ground is not transparent; nor, for that matter, is the sky, to which meteorologists look for information regarding atmospheric and weather patterns. Some places are hard to see, and many are difficult or even impossible to visit physically. Some places, such as the Sun or the Earth's core, could not be approached physically even by unmanned technology.

Hence the need for remote sensing, or the gathering of data without actual contact with the materials or objects being studied. Some earth scientists define the term more narrowly, restricting "remote sensing" to the use of techniques involving radiation on the electromagnetic spectrum. The latter category includes visible, infrared, and ultraviolet light as well as lower-frequency signals in the microwave range of the spectrum. This definition excludes the study of force fields involving gravitational or electromagnetic force. In general, in this essay we abide by that more narrow definition, primarily because most forms of remote sensing in use today involve electromagnetic radiation.

Remote sensing is used for a variety of measuring and mapping applications. The reader therefore is encouraged to consult the essay Measuring and Mapping Earth for more on this subject. Applications of remote sensing go far beyond cartography (mapmaking) and measurement, however. As suggested already, remote sensing makes it possible for earth scientists to collect data from places they could not possibly go. In addition, it allows for data collection in places where a human being would be "unable to see the forest for the trees"—which in places such as the Amazon valley is quite literally the case.

The Military Influence

Scientists' understanding of the electromagnetic spectrum was still in its infancy in 1849, when the French army engineer Aimé Laussedat (1819–1907) introduced what was then called iconometry, from the Greek words *icon* ("image") and *-metry* ("measurement"). Laussedat, who experimented with aerial photography

A hand-held global positioning device. *(© Ken M. Johns/Photo Researchers. Reproduced by permission.)*

by means of cameras mounted on balloons or kites, is regarded as a pioneer of photogrammetry, the use of aerial or satellite photography to provide measurements of or between objects on the ground.

A few years later, the United States armies of the Civil War adopted the use of aerial photography for surveillance purposes, mounting cameras on balloons to provide intelligence regarding federal or Confederate positions and troop strength. This fact, combined with Laussedat's status as an army engineer, hints at one of the underlying themes in the history of remote sensing, and indeed of many another technological advance: the influence of the military. It is a fact of human existence that nations from at least the time of the Assyrians, if not the Egyptians of the New Kingdom, have devoted far more attention and resources to military applications than they have to peacetime activities.

On the other hand, societies have benefited enormously from technological and organizational innovations with military origins, innovations whose application later spread to a variety of peacetime uses. Some examples include the adoption of the chariot by the Egyptian army after the Hyksos invasion (*ca.* 1670 B.C.); the Assyrian introduction of logistics in an effort to supply imperial troops (*ca.* 800 B.C.); the Persian development of the postal service (*ca.* 600 B.C.); numerous Roman innovations, particularly in road building (*ca.* 200 B.C.–*ca.* A.D. 200); and the Chinese invention of the wheelbarrow (*ca.* 100 B.C.). And so the list goes, right up to such latter-day American developments as the Internet and GPS, or global positioning system.

MILITARY CONTRIBUTIONS TO REMOTE SENSING. Forms of technology pioneered by military forces and now used in remote sensing include infrared photography, thermal imagery, radar scanning, and satellites. The first of these types of technology makes use of light in the infrared portion of the electromagnetic spectrum—a region that, as its name suggests, is adjacent to the red portion of visible light. Red has the longest wavelength and the lowest frequency of all colors, and infrared has an even longer wavelength and lower frequency. Military forces use infrared photography to distinguish between vegetation and camouflage designed to look like vegetation: live plants reflect infrared radiation, whereas dead ones and camouflaged material absorb it.

Whereas infrared photography measures reflection of infrared radiation, thermal imaging indicates the amount of such radiation that is emitted by the source. Its military origins lie in its use for reconnaissance during night bombing missions. Similarly, radar scanning makes it possible to view targets on the ground, regardless of lighting or cloud cover. Finally, there are satellites, which have extensive surveillance applications. Among the most important examples of military activity above Earth's atmosphere are the 24 satellites of GPS, which make allow U.S. forces to plot positions with amazing accuracy. Less accurate GPS intelligence is also available to civilians. (See Measuring and Mapping Earth for more on GPS.)

REAL-LIFE APPLICATIONS

Photogeology

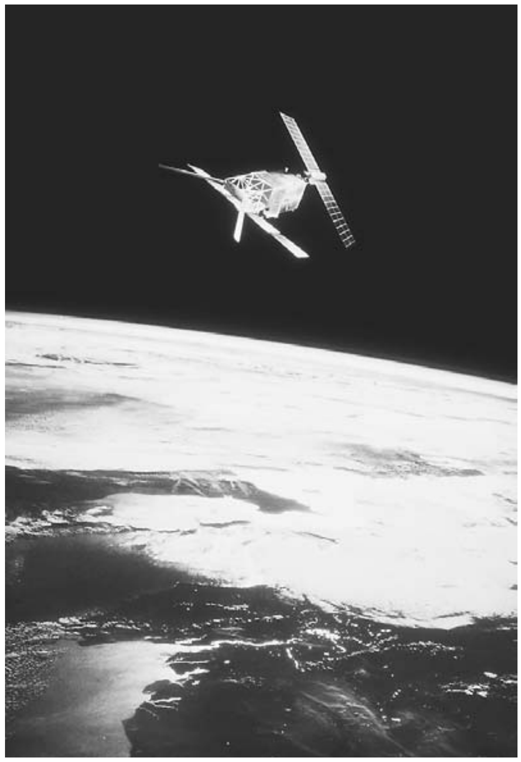
All of these innovations introduced by the military, of course, have found application for civilian purposes. Thanks in part to improvements in aircraft during World War II, for instance, photogeologic data gathering has increased dramatically in the years since then. Efforts at gaining information by means of airborne sensing devices underwent enormous improvements throughout the middle and latter part of the twentieth century, with the development of technology that made it possible for earth scientists to gather information using techniques beyond ordinary photography, visible light, and airplanes.

Still, much of the remote-sensing activity that takes place today is performed aboard airplanes rather than satellites, using ordinary analogue photography within the visible spectrum. Stereoscopic techniques aid in the visualization of relief, or elevation and other inequalities on a land surface. Humans are used to seeing stereoscopically: the distance between the two eyes on our faces results in a difference between the two images each eye sees. The brain corrects for this difference, rendering a stereoscopic image that is more full and dimensional than anything a single eye could produce. The use of multiple cameras and stereoscopic technology replicates this activity of the human brain and thus provides earth scientists with much more information than they could gain simply by looking at "flat" photographs taken from an airplane.

The materials studied by a geologist, of course, are primarily underground, but Earth's surface furnishes many clues that a trained observer can interpret. Uplands and lowlands tend to suggest different types of rocks, while the direction of a dip in the land can supply volumes of information regarding the stratigraphic characteristics of the region. The presence of vegetation can make it harder to discern such clues, but a careful study of plant life can reveal much regarding minerals in the soil, local water resources, and so on.

DIGITAL PHOTOGRAPHY

Within both photogeology and the larger realm of remote sensing, several innovations from the 1960s onward have underpinned more effective methods of observation. One of these is digital photography, which is as much of an improvement over old-fashioned photography as compact discs are over phonograph records. In both cases, the contrast is between analog technology and digital technology. In analog photography,



A COMMUNICATIONS SATELLITE IN ORBIT AROUND EARTH. *(© ESA/Photo Researchers. Reproduced by permission.)*

for instance, the image is recorded by a camera and stored on photosensitive materials in a film emulsion. In digital photography the image is recorded on a solid-state device called an image sensor and stored in the camera's memory for transfer to a computer.

An analogue (the preferred spelling for the word as a noun) is just that, a "close copy," whereas digital methods make possible a more exact reproduction of images by assigning to each shade of color a number between 0 and 255. Instead of storing the image in a medium that can be destroyed or lost easily, as is the case with ordinary film, digital images can be saved on a computer, backed up, and sent anywhere in the world via the Internet. Furthermore, these images can be adjusted with the use of a computer, so as to make it easier to see certain features.

Computers and digital photography aid in the creation of false-color imaging, a means of representing invisible electromagnetic data by assigning specific colors to certain wavelengths. An example would be the use of red to depict areas of high energy. This is certainly a false use of color, since red actually has the lowest energy

AN AERIAL PHOTOGRAPH SHOWS THE COLORADO RIVER DELTA IN THE GULF OF CALIFORNIA. THE RIVER ITSELF IS THE DARK HEMISPHERE AT THE BOTTOM, WITH ITS WATERS BRANCHING OUT THROUGH SANDBARS LIKE THE BOUGHS OF A TREE. *(© Photo Researchers. Reproduced by permission.)*

in the visible spectrum, with purple possessing the highest energy. (The reason we associate red, orange, and yellow with heat and green, blue, and purple with coldness is that in either case, these are the colors objects *reflect,* not the ones they absorb.)

RADAR. Most remote-sensing technology uses light, whether infrared or visible, that falls at the middle to high end of the electromagnetic spectrum. By contrast, at least one important means of remote detection uses microwaves, which are much lower in energy levels. Microwaves carry FM radio and television signals, as well as radar, or *RA*dio *D*etection *A*nd *R*anging.

Radar makes it possible for pilots to "see" through clouds, rain, fog, and all manner of natural phenomena—not least of which is darkness. It also can identify objects, both natural and man-made, on the ground. In addition to its application in remote sensing, radar using the Doppler effect (the change in the observed frequency of a wave when the source of the wave is moving with respect to the observer) helps meteorologists track storms.

In the simplest model of radar operation, a sensing unit sends out microwaves toward the target, and the waves bounce back off the target to the unit. In a monostatic unit—one in which the transmitter and receiver are in the same location—the radar unit has to be switched continually between sending and receiving modes. Clearly, a bistatic unit—one in which the transmitter and receiver antennas are at locations remote from one another—is generally preferable, but on an airplane, for instance, there is no choice but to use a monostatic unit.

Satellite Data

The term satellite refers to any object orbiting a larger one; thus, Earth's Moon and all the other moons of the solar system are satellites, as are the many artificial satellites that orbit Earth. In practice, however, most people use the term to refer only to artificial satellites, of which there are many hundreds, launched by entities ranging from national governments to international associations to independent firms. Artificial satellites typically are intended for the purposes of gathering information (i.e., scientific research or military surveillance) or disseminating it (i.e., through satellite television broadcasting).

In launching a satellite, it is necessary to overcome the enormous pull of Earth's gravitational field. This is done by providing the satellite with power through rocket boosters that launch it far above Earth's atmosphere. At a height of 200 mi. (320 km) or more, the satellite is far above the dense gases of the atmosphere yet well within the gravitational field of the planet. The craft is then in a position to orbit Earth indefinitely without the need for additional power from man-made sources; instead, Earth's own gravitational energy keeps the satellite in orbit for as long as the satellite's structure remains intact. (See Gravity and Geodesy for more about the mechanics of orbit.)

The greater the altitude, the longer it takes a satellite to complete a single revolution. One of the most commonly used altitudes is at 22,500 mi. (36,000 km), at which height a satellite takes 24 hours to orbit Earth. Thus, it is said to be in geosynchronous orbit, meaning that it revolves at the same speed as the planet itself and therefore remains effectively stationary over a given area. Some satellites revolve at even higher altitudes—25,000 mi. (40,225 km), which, while it is far beyond the atmosphere, is well within Earth's gravitational field.

LANDSAT. One of the most impressive undertakings in the field of satellite research is Landsat, an Earth-monitoring satellite designed specifically for the use of earth scientists and resource managers. Conceived by the United States Department of the Interior in the mid-1960s, the Landsat project soon came to involve the National Aeronautics and Space Administration (NASA) and the U.S. Geological Survey (USGS; see Measuring and Mapping Earth for more about geologic surveys.) *Landsat 1* went into orbit on July 23, 1972.

Over the years, Landsat has gone into six subsequent generations. *Landsat 6*, launched in 1993, was unable to achieve orbit, but *Landsat 1* lasted more than five times as long as its projected life expectancy of one year. Since 1972 at least one Landsat satellite has been in orbit over Earth, and as of early 2001 both *Landsat 5* (launched in March 1984) and *Landsat 7* (launched in April 1999) were on line. (*Landsat 5* was decommissioned in June 2001.) Over the course of the years, the Landsat governing body has changed. In the 1980s, NOAA (National Oceanic and Atmospheric Administration) took over from NASA, and in October 1985 the Landsat system came under the direction of a commercial organization, the Earth Observation Satellite Company (EOSat).

In contrast to communication satellites, which tend to maintain geosynchronous orbits, Landsat moves at a much lower altitude and therefore orbits Earth much more quickly. *Landsat 7* takes approximately 99 minutes to orbit the planet, thus making 14 circuits in a 24-hour period. Though it never quite passes over the poles, it covers the rest of Earth in swaths 115 mi. (185 km) wide, meaning that eventually it passes over virtually all other spots on the planet.

SATELLITES AT WORK. Landsat and other satellites, such as France's SPOT (Satellite Positioning and Tracking), provide data for governments, businesses, scientific institutions, and even the general public. Following the September 11, 2001, terrorist bombing of the World Trade Center in New York City, for instance, the SPOT U.S. Web site (<http://www.spot.com>) provided viewers with "Images of Infamy": views of downtown Manhattan before and just a few hours after the bombing.

Data from Landsat has been used to study disasters and potential disasters with particular application to the earth sciences. An example is the area of the tropical rainforest in Brazil's Amazon River valley, a region of about 1.9 million sq.

KEY TERMS

CARTOGRAPHY: The creation, production, and study of maps. Cartography is a subdiscipline of geography and involves not only science but also mathematics, technology, and even art.

DOPPLER EFFECT: The change in the observed frequency of a wave when the source of the wave is moving with respect to the observer.

ELECTROMAGNETIC RADIATION: See Electromagnetic spectrum and Radiation.

ELECTROMAGNETIC SPECTRUM: The complete range of electromagnetic waves on a continuous distribution from a very low range of frequencies and energy levels, with a correspondingly long wavelength, to a very high range of frequencies and energy levels, with a correspondingly short wavelength. Included on the electromagnetic spectrum are long-wave and short-wave radio; microwaves; infrared, visible, and ultraviolet light; x rays; and gamma rays.

FALSE-COLOR IMAGING: A means of representing invisible electromagnetic data by assigning specific colors to certain wavelengths.

FREQUENCY: The number of waves, measured in Hertz, passing through a given point during the interval of one second. The higher the frequency, the shorter the wavelength.

GEODESY: An area of geophysics devoted to the measurement of Earth's shape and gravitational field.

HERTZ: A unit for measuring frequency equal to one cycle per second. High frequencies are expressed in terms of kilohertz (kHz; 10^3, or 1,000 cycles per second), megahertz (MHz; 10^6, or one million cycles per second), and gigahertz (GHz; 10^9, or one billion cycles per second).

PHOTOGEOLOGY: The use of aerial photographic data to make determinations regarding the geologic characteristics of an area.

PHOTOGRAMMETRY: The use of aerial or satellite photography to provide measurements of or between objects on the ground.

RADIATION: The transfer of energy by means of electromagnetic waves, which require no physical medium (for example, water or air) for the transfer. Earth receives the Sun's energy, via the electromagnetic spectrum by means of radiation.

RELIEF: Elevation and other inequalities on a land surface.

REMOTE SENSING: The gathering of data without actual contact with the materials or objects being studied.

STRATIGRAPHY: The study of rock layers, or strata, beneath Earth's surface.

WAVELENGTH: The distance between a crest and the adjacent crest or a trough and the adjacent trough of a wave. Wavelength is inversely related to frequency, meaning that the shorter the wavelength, the higher the frequency.

mi. (five million sq km), in which deforestation is claiming between 4,250 sq. mi. and 10,000 sq. mi. (11,000–26,000 sq km) a year. This is an extremely serious issue, because the Amazon basin represents approximately one-third of the total rainforest area on Earth. Earlier estimates, however, had suggested that deforestation was claiming up to three times as much as it actually is, and Landsat provided a more accurate figure.

Because of its acute spatial resolution (98 ft., or 30 m, compared with more than 0.6 mi., or 1 km), Landsat is much more effective for this purpose than other satellite systems operated by NOAA or other organizations. It is also cheaper to obtain images from it than from SPOT. Over the years, Landsat has provided data on urban sprawl in areas as widely separated as Las Vegas, Nevada, and Santiago, Chile. It has offered glimpses of disasters ranging from the eruption of Mount Saint Helens, Washington, in 1980 to some of the most potent recent examples of destruction caused by humans, including the nuclear disaster at Chernobyl, Ukraine, in 1986 and the fires and other effects of the Persian Gulf War of 1990–1991. (For more on this subject, see the Earthshots Web site, operated by USGS.)

Not all the news from Landsat is bad, as a visit to the *Landsat 7* Web site (<http://landsat.gsfc.nasa.gov/>) in late 2001 revealed. Certainly there were areas of concern, among them, flooding in Mozambique and runaway development in Denver, Colorado. But images taken over the Aldabra atoll in the Seychelles showed the world's largest refuge for giant tortoises. And shots taken from Landsat over Lake Nasser in southern Egypt during the latter part of 2000 showed four lakes created by excess water from Nasser. As a result, that region of the Sahara had new lakes for the first time in 6,000 years.

WHERE TO LEARN MORE

Burtch, Robert. *A Short History of Photogrammetry* (Web site). <http://users.netonecom.net/~rburtch/sure340/history.html>.

"Earthshots: Satellite Images of Environmental Change," U.S. Geological Survey (Web site). <http://edcwww.cr.usgs.gov/earthshots/slow/tableofcontent.

Hancock, Paul L., and Brian J. Skinner. *The Oxford Companion to the Earth.* New York: Oxford University Press, 2000.

NASA EOS IDS Volcanology (Web site). <http://eos.pgd.hawaii.edu/>.

Remote Sensing Data and Information (Web site). <http://rsd.gsfc.nasa.gov/rsd/RemoteSensing.html>.

Skinner, Brian J., Stephen C. Porter, and Daniel B. Botkin. *The Blue Planet: An Introduction to Earth System Science,* 2d ed. New York: John Wiley and Sons, 1999.

Smith, David G. *The Cambridge Encyclopedia of Earth Sciences.* New York: Cambridge University Press, 1981.

Strain, Priscilla, and Frederick Engle. *Looking at Earth.* Atlanta: Turner Publishing, 1992.

Visualization of Remote Sensing Data (Web site). <http://rsd.gsfc.nasa.gov/rsd/>.

The WWW Virtual Library: Remote Sensing (Web site). <http://www.vtt.fi/aut/rs/virtual/>.

PLANETOLOGY

PLANETARY SCIENCE

SUN, MOON, AND EARTH

PLANETARY SCIENCE

CONCEPT

The term *planetary science* encompasses a whole range of studies involving a combination of earth sciences and astronomy. Sometimes known as planetology or planetary studies, these disciplines are concerned primarily with the geologic, geophysical, and geochemical properties of other planets. They also draw on areas of astronomy, such as cosmology, a fascinating discipline devoted to the study of the origin, structure, and evolution of the universe. As always when considering realms beyond our Earth, there are many surprises. Indeed, the more one learns about Earth's relationship to the rest of the cosmos, the harder it is to say which is more intriguing: the many factors that make Earth different or the myriad ways that our home planet is just like the rest of the known universe.

HOW IT WORKS

Earth and the Cosmos

Most of us spend our daily lives without devoting a great deal of thought to what lies beyond Earth. People who live outside cities are perhaps more attuned to the cosmos than are their urban counterparts, simply because they see the vast oceans of stars that cover the sky on a clear night. But a person who lives in the city, where bright lights and smog conspire to cover all but the brightest heavenly bodies, rarely finds a reason to look up into the night sky.

One reason people spend little time thinking about the cosmos is that to do so ultimately fills one with a sense of awe bordering on dread. We know that Earth is but one planet of nine, revolving around an average-sized star, the Sun, somewhere between the center and the edge of a galaxy called the Milky Way—itself just one of many galaxies in the universe. This awareness naturally makes a person feel small and almost inevitably raises questions about the nature of the soul, divinity, and the afterlife.

RELIGION, PHILOSOPHY, MYSTICISM, AND SCIENCE. Such questions are a natural accompaniment to of our feeling that if one person is so truly insignificant in this vast cosmos, there must be something else that gives meaning to the structure of reality. These vast issues, of course, are properly addressed not by science but by theology and philosophy. Science, on the other hand, is concerned simply with the facts of how the universe emerged and how Earth fits into the larger picture.

Yet it is easy to see how ancient peoples would have perceived no distinction between religion and science where the study of the cosmos was concerned. The Babylonians, for instance, had no concept of any difference between scientific astronomy and astrology, which today is recognized as a superstitious and thoroughly unscientific pursuit. The Greeks modeled the cosmos on their philosophical systems, which provided a hierarchy of material forms and an ordered arrangement of causes and substances. And the Judeo-Christian tradition depicts a universe fashioned by a loving, all-powerful creator who designed the human being in his own image.

In the belief systems of Judaism and Christianity, handed down through the Bible, the cosmos is depicted as the setting of a vast spiritual drama centered around the themes of free will, sin, and redemption. The Bible never says that Earth is the center of the physical universe, but it clearly presents it as the center of the spiritual one. This is understandable enough, especially if human beings truly are the only intelligent life-forms; unfortunately, these spiritual ideas eventually informed an erroneous cosmology that depicted Earth as the physical center of the cosmos.

COSMOLOGY

In fairness to Christianity, it should be said that most religious, philosophical, and even scientific traditions before about 1500 depicted Earth as the center of the universe. Indeed, it required a great feat of insight to discern that Earth is not the center. The same is true of many other discoveries about the cosmos, where nothing is as it appears when simply gazing into the night sky.

In a scene from his great novel *The Adventures of Huckleberry Finn* (1884), Mark Twain aptly illustrated the impossibility of understanding the universe simply on the basis of unaided intellect. Huck and the runaway slave Jim have just finished supper and are lying on their backs and staring up at the stars, speculating as to their origins. One of them comes up with a theory that seems altogether plausible on the face of it: the Moon, because it looks larger than the stars, must have laid them like eggs. A similar scene occurs in the children's movie *The Lion King* (1994), in which one character postulates that the stars have become stuck to the sky like flies on flypaper. When another character, the warthog Pumbaa, correctly suggests that the stars are actually great balls of burning gas billions of miles away, his companions laugh this off as preposterous.

ARISTARCHUS AND HIPPARCHUS. Although they lacked telescopes, the Greeks developed rather sophisticated (though in many cases wrong) ideas concerning the arrangement of the cosmos. Most notable among these early thinkers was the astronomer Aristarchus of Samos (*ca.* 320–*ca.* 250 B.C.), who proposed that Earth rotates on its axis once every day and revolves with other planets around the Sun. He also correctly suggested that the Sun is larger than Earth.

Unfortunately, the astronomer Hipparchus (146–127 B.C.) rejected this heliocentric, or Sun-centered, cosmology in favor of a geocentric, or Earth-centered, model. Among Hipparchus's later followers was the Alexandrian Ptolemy (*ca.* A.D. 100–170), destined to become the most influential astronomer of ancient and medieval times, who established geocentric cosmology as a guiding principle of astronomy.

THE PTOLEMAIC SYSTEM. The influence of Ptolemy's erroneous ideas is partly an accident of history. He lived, as it turned out, in the last great era of civilization: ten years after his death came that of the Roman emperor Marcus Aurelius (A.D. 121–180), whose passing marked the beginning of Rome's decline over the next three centuries. Learning in western Europe virtually ceased until about 1200, and even though the Muslim world produced several thinkers of note during this period, most of them worked within the tradition established by Ptolemy. Muslim thinkers' respect for Ptolemy is reflected in the name that Arab translators gave to his most important writing: *al-majisti* or "majesty." When this work made its way to Europe, it became known as the *Almagest.*

The Ptolemaic system proves that it is possible to prove anything, if one creates a methodology elaborate enough. Of course, as we know now, Earth is not the center of the universe, but pure observation alone did not reveal this, and Ptolemy's cosmology worked because he developed mathematics and ideas of planetary motion that made it workable. For instance, not only did planets orbit around Earth in Ptolemy's cosmology, but they also moved in circles around the paths of their own orbits. Of course, they do revolve on their axes, but that was not part of Ptolemy's model. In fact, it is hard to find an analogy in the real world, with the exception of some bizarre amusement park ride, for the form of motion Ptolemy was describing.

He was trying to explain retrograde motion, or the fact that other planets seem to speed up and slow down. Retrograde motion makes perfect sense once one understands that Earth is moving even as the other planets are moving, thus creating the optical illusion that the others are changing speeds. Since the Ptolemaic system depicted a still Earth in the middle of a moving universe, however, the explanation of retrograde motion required mental acrobatics.

CHALLENGING PTOLEMY. Although it is incorrect, the Ptolemaic system was a creation of genius; otherwise, it could not have survived for as long as it did. Even with the recovery of learning in Europe during the late Middle Ages, scientists continued to uphold Ptolemy's ideas. Instead of discarding his system, or at least calling it into question, astronomers simply adjusted the mathematics and refined their ancient forebear's physical model to account for any anomalies.

The revolution against Ptolemy began quietly enough in the fifteenth century, when the Austrian astronomer and mathematician Georg Purbach (1423–1461) noted the inaccuracies of existing astronomical tables and the need for better translations of Greek texts. Purbach attempted to produce a revised and corrected version of the *Almagest,* but he died before completing it. The job fell to his student, Johann Müller, who was known as Regiomontanus (1436–1476).

The *Epitome of the Almagest* (1463), begun by Purbach and completed by Regiomontanus, proved to be a turning point in astronomy. Like their medieval predecessors, the two men started out working in the Ptolemaic tradition, but by showing the errors in Ptolemy's work, they actually were criticizing him. Their discoveries were not lost on a young Polish astronomer named Nicolaus Copernicus (1473–1543).

THE COPERNICAN REVOLUTION. The story of the Copernican Revolution, the opening chapter in a larger movement known as the Scientific Revolution, is among the greatest sagas in the history of thought. It was a watershed event, marking the birth of modern science as such, but the change in thought patterns created by this revolution was not so much the work of Copernicus as it was of the Italian astronomer Galileo Galilei (1564–1642). Although he often is given less attention than Copernicus and the other most noted figure of the Scientific Revolution, the English natural philosopher Isaac Newton (1642–1727), Galileo was a thinker of the first order who took Copernicus's discoveries much further.

Copernicus had been concerned with how the planets move as they do, and in the course of his work he showed that all of them (Earth included) move around the Sun. Galileo, on the other hand, set out to discover *why* the planets revolve around the Sun, and in so doing he discovered the principles of inertia and gravitational acceleration that would influence Newton. He made numerous other contributions, such as the discovery that Jupiter had moons, but by far his greatest gift to science was his introduction of the scientific method.

Thanks to Galileo and others who later refined the method, thinkers would no longer be content to let mere conjecture guide their work. Before his time, scientists generally had followed a pattern of absorbing the received wisdom of the ancients and then seeking evidence that confirmed those suppositions. The new scientific method, on the other hand, required rigorous work: detailed observation, the formation of hypotheses, testing of hypotheses, formation of theories, testing of theories, formation of laws, testing of laws—and always more observation and testing.

REAL-LIFE APPLICATIONS

Gravity, the Sun, and Earth

A fourth key figure in the Scientific Revolution was the German astronomer Johannes Kepler (1571–1630), whose laws of planetary motion directly influenced Newton's laws of gravitation and motion. Thanks to Kepler, we know that planets do not make circular orbits around the Sun; rather, those orbits are elliptical. As Newton later showed, the reason for this is the gravitational pull exerted by the Sun.

Gravitational force explains why Earth, the Sun, and all celestial bodies larger than asteroids are round—but also why they cannot be *perfectly* round. As to the latter issue—the fact that Earth bulges near the equator—it is a consequence of its motion around its axis. Because it is spinning rapidly, the mass of the planet's interior responds to the centripetal (inward) force of its motion, producing a centrifugal, or outward, component. If Earth were standing still, it would be much nearer to the shape of a sphere.

MASS AND SPHERICITY. Now to the larger question: Why is Earth round? The answer is that the gravitational pull of its interior forces a planetary body to assume a more or less uniform shape. Furthermore, the larger the mass of an object, the greater its tendency toward roundness. Earth's surface has a relatively small vertical differential: between the lowest point and the highest point is just 12.28 mi. (19.6 km), which is not a great distance, considering that Earth's radius is about 4,000 mi. (6,400 km).

JOHANNES KEPLER *(The Bettmann Archive. Reproduced by permission.)*

An object of less mass is more likely to retain a shape that is less than spherical. This can be shown by reference to the Martian moons Phobos and Deimos, both of which are oblong, and both of which are tiny, in terms of size and mass, compared with Earth's Moon. Mars itself has a radius half that of Earth, yet its mass is only about 10% of Earth's, and therefore it is capable of retaining a less perfectly spherical shape.

There is also the possibility of more pronounced differences in elevation, and thus it should not be surprising to learn that Mars is also home to the tallest mountain in the solar system. Standing 15 mi. (24 km) high, the volcano Olympus Mons is not only much taller than Earth's tallest peak, Mount Everest (29,028 ft., or 8,848 m), it is also 22% taller than the distance from the top of Mount Everest to the lowest spot on Earth, the Mariana Trench in the Pacific Ocean (–36,198 ft., or –10,911 m).

WHY EARTH IS SPECIAL

With regard to gravitation, a spherical object behaves as though its mass were concentrated near its center. Indeed, 33% of Earth's mass is at is core (as opposed to the crust or mantle), even though the core accounts for only about 20% of the planet's volume. Geologists believe that the composition of Earth's core must be molten iron, which creates the planet's vast electromagnetic field.

Certain particulars of Earth's core lead us to answering another great question about our home planet: Why is it alone capable of sustaining life—as far as we can tell—while the other planets of our solar system are either hellish worlds of fire or frigid, forbidding realms of ice crystals and liquefied gas?

DENSITY OF EARTH'S INTERIOR. At first glance, Earth seems to have few distinctions other than its ability to support life: it is neither the largest nor the smallest planet in the solar system, positions held by Jupiter and Pluto, respectively. (Earth ranks fifth.) Earth has a moon, but that is hardly a distinction: Saturn has 18 moons. And not only does Olympus Mons tower over Everest, but the gaseous oceans of Jupiter also are much deeper than the Mariana Trench. In the lists of planetary superlatives, Earth has only one: it is the most dense.

The only bodies that come close are Mercury and Venus, which along with Earth and Mars are designated as terrestrial planets. (Earth's Moon often is considered along with the terrestrial planets because its composition is similar to them and because it is a relatively large satellite.) The terrestrial planets are small, rocky, and dense; have relatively small amounts of gaseous elements; and are composed primarily of metals and silicates. This is in contrast to the Jovian planets, which are large, low in density, and composed primarily of gases. (The Jovian planets usually are designated as the four giants Jupiter, Saturn, Uranus, and Neptune. Pluto, the smallest of all nine planets, has a density higher than any Jovian planet.)

Density is simply the ratio of mass to volume, meaning that Earth packs more mass into a given volume than any other body in the solar system. Saturn, least dense among the planets, has a mass 95.16 times as great as that of Earth, yet its volume is 764 times greater, meaning that its density is only about 12% of Earth's. But whereas Saturn and other Jovian planets are composed primarily of gases surrounding small, dense cores, Earth—beneath its atmospheric layer and it waters—is a hard little ball. Its core, composed of iron, nickel, and traces of other elements, including uranium, is relatively heavy.

That gives it a strong gravitational pull and, in combination with the comparatively high speed of the planet's rotation, causes Earth to have a powerful magnetic field. It is also important to note the significance of planetary mass in making possible the formation of an atmosphere. Because of their mass, larger planetary bodies exert enough gravitational pull to retain gases around their surfaces; by contrast, the Moon and Mercury are too small and have no atmosphere. Of course, Earth is the only planet whose atmosphere is capable of sustaining life as we know it, and this is a result of activity beneath the planet's surface.

A VOLATILE PLANET. Earth is the only terrestrial planet on which the processes of plate tectonics, or the shifting of plates beneath the planetary surface, take place. The other terrestrial planets have crusts of fairly uniform thickness, suggesting that they have never experienced the internal shifting that has helped give our planet its unique topography. Earth also has a relatively thin lithosphere—the upper layer of the planetary surface, including the crust and the brittle portion at the top of the mantle—which helps make it a particularly volatile body.

Of the terrestrial planets, the only ones still given to volcanic activity are Earth and Venus. Mars seems to have experienced volcanic activity at some point in the past billion years, while Mercury and the Moon have not had volcanoes for several billion years. This is also an important factor in determining Earth's capacity to support living things, because volcanoes—which transport gases from the planet's interior to its atmosphere—have been crucial to the creation of the conditions necessary for sustaining life.

The heat generated by internal volatility is also a component influencing the sustainability of life on Earth. At the time Earth and other planets were formed, some 4.5 billion years ago, the planets experienced such heat that they melted, causing a separation of chemical compounds. The heavier compounds, mostly containing iron, sank to the core of the planet, where they remain today, while the lighter ones rose to the surface. Included in these lighter substances were oxygen and other elements essential to the sustenance of life. Even now Earth and Venus, because of their volcanic activity, are cooling at rates slower than those of the other planets, and this has facilitated the separation of elements.

THE SILHOUETTE OF A STARGAZER AGAINST THE MILKY WAY. *(© F. Zullo/Photo Researchers. Reproduced by permission.)*

THE CREATION AND SUSTENANCE OF LIFE. Aside from the distinctive features of its core, Earth's position relative to the Sun has helped make it possible for life to take root on this planet. For decades scientists believed that Earth is unique in possessing that life-sustaining compound of hydrogen and oxygen, H_2O or water; but now we know that even Jupiter—not to mention Venus and Mars—have water on their surfaces. The problem is that Venus's water is too hot, existing as vapor in the upper atmosphere, while the water on Mars and Jupiter takes the form of ice crystals. Earth is uniquely placed to sustain liquid water.

The existence of liquid water made it possible for the first microorganisms to form on Earth, leading over hundreds of millions of years to the development of the complex biosphere known today. The existence of life in simple forms promoted the development of the atmosphere and geosphere, because these life-forms took in carbon dioxide and water, processed them, and returned them to the environment as oxygen and organic materials.

COMPUTER IMAGE OF THE NINE PLANETS OF OUR SOLAR SYSTEM. *(© Photo Researchers. Reproduced by permission.)*

THE SOLAR SYSTEM AND BEYOND

The reader may have noticed that earlier in this essay, we ceased discussing progress in cosmology after about 1650. This is not because nothing happened after that time; on the contrary, the centuries that have elapsed since then have seen the greatest progress in astronomical study since the dawn of civilization. To give this topic the coverage it warrants, however, would require a lengthy discussion—one that would take us away from the earth sciences and toward the sister science of astronomy.

Up until the Scientific Revolution, the earth sciences hardly existed, except inasmuch as various people over the millennia had recorded data concerning Earth and made sometimes unscientific speculations regarding its origin and composition. As the oldest of the physical sciences, astronomy was much more mature, but even it could progress only so far under the restrictions of the Ptolemaic system. Unfettered, it began to progress rapidly, and the result has been an unfolding vision of the universe that is at once more clear and more complex.

THE SIZE OF THE UNIVERSE. One of the dominant themes in astronomy from Galileo's time to the present day is astronomers' quite literally expanding vision of the universe. Up until 1781, when the German-born English astronomer William Herschel (1738–1822) discovered Uranus, scientists had known only of the

KEY TERMS

ATMOSPHERE: In general, an atmosphere is a blanket of gases surrounding a planet. Unless otherwise identified, however, the term refers to the atmosphere of Earth, which consists of nitrogen (78%), oxygen (21%), argon (0.93%), and other substances that include water vapor, carbon dioxide, ozone, and noble gases such as neon, which together comprise 0.07%.

BIOSPHERE: A combination of all living things on Earth—plants, mammals, birds, reptiles, amphibians, marine life, insects, viruses, single-cell organisms, and so on—as well as all formerly living things that have not yet decomposed. Typically, after decomposing, a formerly living organism becomes part of the geosphere.

CENTRIFUGAL: A term describing the tendency of objects in uniform circular motion to move outward, away from the center of the circle. Though the term *centrifugal force* often is used, it is inertia, rather than force, that causes the object to move outward.

CENTRIPETAL FORCE: The force that causes an object in uniform circular motion to move inward, toward the center of the circle.

COMPOUND: A substance made up of atoms of more than one element chemically bonded to one another.

COSMOLOGY: The study of the origin, structure, and evolution of the universe.

COSMOS: The universe.

ELECTROMAGNETIC ENERGY: A form of energy with electric and magnetic components that travels in waves and which, depending on the frequency and energy level, can take the form of long-wave and short-wave radio; microwaves; infrared, visible, and ultraviolet light; x rays, and gamma rays.

ELEMENT: A substance made up of only one kind of atom. Unlike compounds, elements cannot be chemically broken into other substances.

GEOCENTRIC: Earth-centered.

GEOCHEMISTRY: A branch of the earth sciences, combining aspects of geology and chemistry, that is concerned with the chemical properties and processes of Earth.

GEOLOGY: The study of the solid earth, in particular, its rocks, minerals, fossils, and land formations.

GEOPHYSICS: A branch of the earth sciences that combines aspects of geology and physics. Geophysics addresses the planet's physical processes as well as its magnetic and electric properties and the means by which energy is transmitted through its interior.

GEOSPHERE: The upper part of Earth's continental crust, or that portion of the solid earth on which human beings live and which provides them with most of their food and natural resources.

HELIOCENTRIC: Sun-centered.

HYDROSPHERE: The entirety of Earth's water, excluding water vapor in the atmosphere but including all oceans, lakes, streams, groundwater, snow, and ice.

HYPOTHESIS: An unproven statement regarding an observed phenomenon.

KEY TERMS CONTINUED

INERTIA: The tendency of an object in motion to remain in motion and of an object at rest to remain at rest.

JOVIAN PLANETS: The planets between Mars (the last terrestrial planet) and Pluto, all of which are large, low in density, and composed primarily of gases.

LAW: A scientific principle that is shown always to be the case and for which no exceptions are deemed possible.

LITHOSPHERE: The upper layer of Earth's interior, including the crust and the brittle portion at the top of the mantle.

MANTLE: The layer, approximately 1,429 mi. (2,300 km) thick, between Earth's crust and its core.

MASS: A measure of inertia, indicating the resistance of an object to a change in its motion. (By contrast, weight—which people tend to think of as similar to mass—is a measure of gravitational force, or mass multiplied by the acceleration due to gravity.)

PLANETARY SCIENCE: The branch of the earth sciences, sometimes known as planetology or planetary studies, that focuses on the study of other planetary bodies. This discipline, or set of disciplines, is concerned with the geologic, geophysical, and geochemical properties of other planets but also draws on aspects of astronomy, such as cosmology.

PROTON: A positively charged particle in the nucleus of an atom.

SCIENTIFIC METHOD: A set of principles and procedures for systematic study that includes observation; the formation of hypotheses, theories, and laws; and continual testing and reexamination.

SCIENTIFIC REVOLUTION: A period of accelerated scientific discovery that completely reshaped the world. Usually dated from about 1550 to 1700, the Scientific Revolution saw the origination of the scientific method and the introduction of such ideas as the heliocentric (Sun-centered) universe and gravity.

TERRESTRIAL PLANETS: The four inner planets of the solar system: Mercury, Venus, Earth, and Mars. These planets are all small, rocky, and dense; have relatively modest amounts of gaseous elements; and are composed primarily of metals and silicates. Compare with Jovian planets.

THEORY: A general statement derived from a hypothesis that has withstood sufficient testing.

UNIFORM CIRCULAR MOTION: The motion of an object around the center of a circle in such a manner that speed is constant or unchanging.

five other planets visible to the naked eye, all of which had been discovered in prehistoric times. (Neptune was discovered in 1846 and Pluto not until 1930.)

In the seventeenth century, astronomers still regarded what we call the solar system as the entire universe, but Herschel was instrumental in ascertaining that Earth is part of a bright band of stars called the Milky Way. Just as Earth once had been believed to be the center of the "universe," or solar system, astronomers then came to believe it was at the center of the Milky Way. Only since 1920 has it been known that our solar system is, in fact, somewhere between the center and the edge of the vast galaxy. Even the Milky Way, composed of several hundred billion stars and about 120,000 light-years in diameter, is not the entire universe; it is only one of many hundreds of galaxies or "island universes."

As discussed at the beginning of this essay, such a scale is almost too much for the human mind to comprehend, particularly inasmuch as Earth is the only planet known to sustain intelligent life. As the British science-fiction writer Arthur C. Clarke (1917–) has observed, either there are other intelligent life-forms out there in the universe, or there are not—and either possibility is mind-boggling.

THE BIG BANG AND THE SOLAR SYSTEM. Not only has astronomers' understanding of the universe expanded, along with their idea of its size; it also appears that the universe itself is expanding. Today the most widely accepted model regarding the formation of the universe is the big bang theory, first put forward by the Belgian astrophysicist Georges Édouard Lemaître (1894–1966) in 1927. According to this theory, an explosion 10-20 billion years ago resulted in the rapid creation of all matter in the universe, and that matter is continuing to move outward, expanding the frontiers of the universe.

Our own solar system appears to be about five billion years old, meaning that the Sun is a relatively young star. It seems that the future solar system was just one of many great balls of gas, rotating as they moved outward, that were scattered around the universe as a result of the big bang. Just as these balls of gas exploded from the center, the material of the various stars emerged from the center of the ball that became our solar system.

FORMATION OF THE PLANETS. The proto–solar system we have described here was a great rotating cloud, and though it has long since ceased to be a cloud, it continues to rotate—only now it is in the form of planets turning around a sun at the center. The hottest portion of the cloud, at the center, became the Sun, while cooler portions at the fringes became planets. The Sun itself is composed primarily of hydrogen and helium, the two most plentiful elements in the universe. In the extraordinarily high temperatures on the Sun, atoms of hydrogen (which has one proton in its nucleus) experience nuclear fusion, becoming atoms of helium, which has two protons. It appears that continued fusion resulted in the creation of the heavier elements (for instance, nitrogen, carbon, oxygen, and silicon) of which the planets—in particular, our own—are composed.

Earth's elemental makeup is discussed elsewhere in this book, as is the structure of its interior. So, too, is the Sun's effect on Earth. These matters are not unrelated. In studying the solar system and the planets that make it up, one is confronted again and again with the fact that a planet's destiny is governed by its position relative to the Sun. Ultimately, the planets in our solar system are ruled by the same principle that drives the sale of real estate: location, location, location!

This is true not only of the atmosphere and temperature of planets but also of their relative density. It is no mistake that the terrestrial planets are closer to the Sun: their internal composition is as it is because these bodies became the destination of most of the heavier elements that emanated from it. Many of the lighter elements continued to move outward, where they gathered around rocky centers to become the mostly gaseous Jovian planets.

WHERE TO LEARN MORE

Astronomy and Cosmology (Web site). <http://www.netlabs.net/hp/tremor/>.

Beatty, J. Kelly. *The New Solar System.* New York: Sky, 1999.

Cambridge Cosmology (Web site). <http://www.damtp.cam.ac.uk/user/gr/public/cos_home.html>.

Canup, Robin M., and Kevin Righter. *Origin of the Earth and Moon.* Tucson: University of Arizona Press, 2000.

The Cosmology: Explore the Largest Mystery (Web site). <http://library.thinkquest.org/28181/>.

Gallant, Roy A. *Earth's Place in Space.* New York: Benchmark Books, 1999.

Lambert, David, and Martin Redfern. *The Kingfisher Book of the Universe.* New York: Kingfisher, 2001.

Llamas Ruiz, Andrés. *The Origin of the Universe.* Illus. Luis Rizo. New York: Sterling Publishers, 1997.

Ned Wright's Cosmology Tutorial (Web site). <http://www.astro.ucla.edu/~wright/cosmolog.htm>.

Skinner, Brian J., Stephen C. Porter, and Daniel B. Botkin. *The Blue Planet: An Introduction to Earth System Science.* 2d ed. New York: John Wiley and Sons, 1999.

SUN, MOON, AND EARTH

CONCEPT

Earth is intimately tied to the star around which it revolves, the Sun, and the satellite that revolves around Earth itself, the Moon. Without the Sun, of course, life on Earth simply could not exist, not just because of the need for light but to an even greater degree because of the energy it supplies. For that matter, Earth itself would not exist: our planet appears to have developed from the same cosmic cloud that formed the Sun, and without the Sun to hold it in place with its gravitational pull, Earth would go spinning off into space. For all its influence on human life, however, the Sun has less impact on the tides than the Moon, which is smaller but much closer. Together, these two bodies literally define time in human experience, which has been marked by the movements of the Sun and Moon from a time before civilization began.

HOW IT WORKS

Formation of the Sun and Planets

Scientists today believe that the universe began between 10 and 20 billion years ago, with an event nicknamed the "big bang," The galaxies continued (and still continue) to move outward from that center, and among them was our Milky Way. Somewhere between the center and the rim of the Milky Way, a rotating cloud of cosmic gas formed about six billion years ago,

The center of the cloud, where the greatest amount of gases gathered, was naturally the densest and most massive portion as well as the hottest. There, hydrogen—the lightest of all elements—experienced extraordinary amounts of compression, owing to the density of the clouded gases around it, and underwent nuclear fusion, or the bonding of atomic nuclei. This hot center became the Sun about five billion years ago, but there remained a vast nebula of gas surrounding it. As the fringes of this nebula began to cool, the gases condensed, forming solids around which particles began to accumulate. These were the future planets.

THE PLANETS AND THE ELEMENTS. Closest to the Sun, the planets and other satellites were formed of elements that could condense at high temperatures: iron, silicon, aluminum, calcium, and magnesium. These elements, along with oxygen-containing compounds, constituted the material foundation around which other particles accumulated to form the terrestrial planets: Mercury, Venus, Earth, and Mars. Further from the Sun, where temperatures were lower, gaseous compounds—including methane, ammonia, and even water—could condense. These compounds became the basis around which the five outer planets (the four Jovian planets and Pluto) formed.

The process of planetary formation involved additional steps and took place over a period of about 500 million years. Some of the factors that played a part in forming Earth are discussed in Planetary Science, but in the present context, let us consider the source of those elements mentioned here. Did they just magically form? Were they always there? In answer to the first question, there is nothing magical about the formation of "new" elements, though it almost seems so. As for the second question, the answer is yes and no.

Earth with the Moon in eclipse. *(© Chris Bjornberg/Photo Researchers. Reproduced by permission.)*

Elements and the Sun

The elements themselves were not always there in the universe or on Earth itself; however, the subatomic building blocks that make them up have indeed existed from the beginning of the universe. The basic atomic structure is as follows. There is a nucleus in which one or more protons (positively charged subatomic particles) may reside along with one or more neutrons, which have no charge. Spinning around the nucleus are one or more electrons, or negatively charged subatomic particles.

The number of protons and electrons in an atom is always the same, meaning that the atom has no electric charge. Atoms that have lost or gained electrons (in which case they would acquire a positive or negative charge, respectively) are called ions. Electrons, which move fast and possess very small mass compared with protons and neutrons, are very easy to dislodge from an atom; on the other hand, it takes an extraordinary event to change the number of protons in the nucleus.

This fact is significant, because it points toward the defining characteristic of an element: the number of protons in the nucleus. Atoms of a particular element *always* have the same number of protons, called their atomic number. The atomic number of an element can be determined by consulting the periodic table of elements: for instance, iron, with an atomic number of 26, *must* have 26 protons in its nucleus. If an atom has 25, it is manganese, and if it is has 27, it is cobalt.

FORMING NEW ELEMENTS. The number of neutrons in the nucleus may vary for atoms within a given element. Atoms that have the same number of protons (and are thus of the same element) but differ in their number of neutrons are called isotopes. Most isotopes are stable, meaning that their chemical composition will remain as it is; however, some isotopes are radioactive, meaning that they experience the spontaneous emission of particles or energy over a given period of time.

Radioactive decay is one of two ways that one element can become another. When a radioactive isotope emits an alpha particle, for example, its nucleus expels a positively charged nucleus consisting of two protons and two neutrons, which is the same thing as a helium atom stripped of its electrons. This obviously changes the number of protons in the nucleus of the isotope and may result in its stabilization. The other means of forming a new element is by nuclear fusion, in which two atomic nuclei fuse or bond.

NUCLEAR FUSION. Note that the first of these means by which elements are formed is subtractive; in other words, with radioactive decay, a different element is formed by the expulsion of protons. Nuclear fusion, on the other hand, is additive, resulting in the creation of different elements by the addition of protons to an atomic nucleus. Radioactive decay takes place inside Earth (among other places), while nuclear fusion is the source of the Sun's power.

Nuclear fusion involves the release of huge amounts of energy. On Earth, scientists have been able to bring about uncontrolled nuclear fusion in the form of the so-called hydrogen bomb, which is actually a "fusion bomb." They have yet to succeed in creating controlled nuclear fusion. If and when they do, it would provide a safe, clean source of almost limitless power and probably would constitute the greatest scientific or technological discovery since fire.

On the Sun, nuclear fusion has been taking place, and will continue to do so, for a long, long time. The 92 naturally occurring elements of the universe are the result of fusion reactions, meaning that all that we see around us was once part of a star. This represents a major break with the ancient belief that Earth is made of fundamentally different substances than are the bodies of space (see Earth, Science, and Nonscience). In fact, our world and everything in it—including our own bodies—is truly "the stuff of stars."

The Moon and Earth

In comparison to the Sun, the Moon is altogether less remarkable. Below, we review some statistics about the sizes of each, but as every elementary-school student today knows, the Sun is much, much larger and exerts far more impact on the fate of the solar system—including Earth. The Moon does not even have its own energy sources: its light comes from the Sun, and the absence of an atmosphere, of volcanic activity, or even of a significant magnetic field makes it a very dull place indeed.

Yet the Moon has inspired at least as much fascination among humans over the ages as has the Sun. There is its physical beauty, though comparisons with the Sun are hardly fair, since we cannot look at the Sun without damage to our eyes. There is also its influence on earthly cycles ranging from the tides to the months themselves, though some claims of lunar influence have little basis in fact. During the Middle Ages, many believed that the Moon caused madness, a superstition still reflected in our word lunacy.

Still, humans have long associated a spirit of mystery with the Moon, in part because of its ever-changing appearance and in part because it has always showed just one side to Earth. (We discuss the reason why later.) Only in 1959, when the Soviet space probe *Luna* traveled to the "dark side of the Moon," did scientists gain a glimpse of it. Unmanned and later manned journeys, which culminated with the U.S. Moon landing in 1969, also changed astronomers' understanding of the Moon's origins.

THE "BIG WHACK." One of the curious things about the Moon is its size in relation to Earth. Nowhere in the universe is there such a small size differential between a satellite and the planet around which it orbits, the only possible exception being Pluto and its moon, Charon. Because our Moon is so close in size to Earth, scientists once speculated that they might have shared origins, and this speculation informed several theories concerning the formation of the Moon.

According to the fission theory, the Moon was a piece of Earth that had been torn away, perhaps from the Pacific basin. The simultaneous creation theory likewise depicted Earth and the Moon as sharing origins, but in this case they literally had been formed together from the same materials. Finally, there was the capture theory, which, in contrast to the others, assumed quite different origins for the two bodies: the Moon had formed somewhere else in the solar system and had been captured by Earth's gravitational field after it wandered too close to the planet.

As it turned out, the capture theory was closest to the theory accepted today, though it was discarded along with the other two on the basis of data brought back from the Apollo Moon landings. According to the giant impact theory, sometimes called the Big Whack model or the ring ejection theory, at a young age Earth was sideswiped by a celestial object as large or larger than Mars. As a result of that collision, a ring of crustal matter was spewed into space, and over time the matter in this ring agglomerated to form the Moon.

Vital Statistics of the Sun and Moon

The Moon is about 240,000 mi. (385,000 km) from Earth, meaning that it takes about 1.25 seconds for its light to reach Earth. By contrast, the Sun's distance from Earth is so great that it takes eight minutes for sunlight to reach our planet, even though light travels through space at the speed of about 186,000 mi. (299,339 km) per second.

The distance between Earth and the Sun is the basis for the astronomical unit (AU), a figure used for measuring the distance between bodies in the solar system. Equal to the average distance from Earth's center to the center of the Sun, an AU is designated $1.49597870691 \times 10^8$ km, or approximately 92,955,807 mi. Usually we think of the solar system as the area encompassed by the orbit of the most remote planet, Pluto, but that is only 39.44 AU, a tiny figure compared with the diameter of the realm within the Sun's gravitational pull, which is a staggering 100,000 AU.

VOLUME AND MASS. The Sun itself has a diameter of about 856,000 mi. (1,392,000 km), meaning that the distance across it is about 109 times that of Earth's diameter. Another way to consider that figure is this: if one were to draw a circle as big as the Sun around Earth, the edge of that circle would be about twice as far away as the Moon. The Sun's volume is so great that about 1.3 million Earths could fit inside it, and its mass is about 300,000 times that of Earth. In fact, it accounts for about 99.8% of the mass of the entire solar system.

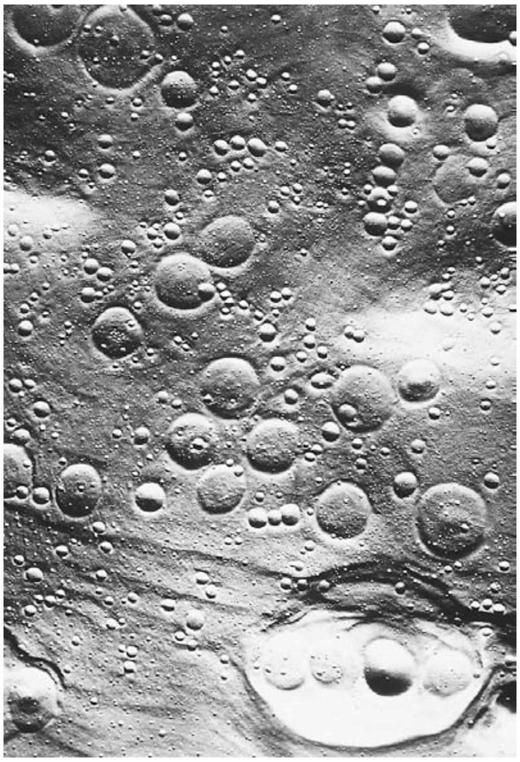
By contrast, the Moon has a diameter of only about 2,160 mi. (3,475 km), a little less than the distance from New York to Los Angeles. Its mass is a little more than 1% of Earth's, and as a result, its gravitational pull is too small to retain the gases that make up the atmosphere. That small mass, combined with an imbalance in its distribution, explains why the Moon shows only one face to Earth. The side of the Moon facing Earth is of greater mass than the other side and is therefore more strongly attracted by Earth's gravitational force. The result is a phenomenon called gravitational locking, whereby the Moon rotates on its axis at exactly the same rate as it travels around Earth—once ever 29.5 days.

Clearly, the Moon is tiny in both volume and mass, and were it as far away as the Sun, we would hardly pay it notice. Yet it should be pointed that even the Sun itself, while remarkable in our own solar system, is far from a standout in the universe as a whole. It is a youngish star, of average size, and not all that different from billions of other stars throughout the cosmos. It is not even unique in being the only star with its own solar system. In 1999 astronomers discovered an entire solar system some 44 light-years from Earth, in which three large planets were found to be circling the star Upsilon Andromedae.

REAL-LIFE APPLICATIONS

Energy, Temperature, and Composition of the Moon

Of course, the area in which the Sun and Moon differ most significantly is in terms of the energy they possess, which, in turn, affects both their



THE MOON'S SURFACE IS POCKMARKED WITH CRATERS, SHOWING ITS VULNERABILITY TO DAMAGE CAUSED BY PARTICLES FROM SPACE. *(© NASA/Photo Researchers. Reproduced by permission.)*

temperature and the composition of each body. With regard to the Moon's temperature, little need be said, since it is entirely a function of the Moon's exposure (or lack of exposure) to heat from the Sun. On the Moon figures range from 280°F (138°C) to –148°F (–100°C), with a mean temperature of –10°F (–23.33°C).

As for the Moon's composition, it has similarities to and differences from that of Earth, and both these similarities and differences are instructive. The internal composition of the Moon is such that it often is treated along with the terrestrial planets. In addition, it is not much smaller than the smallest terrestrial planet, Mercury. But whereas Mercury has a proportionally large, extremely hot core, the Moon's core is much smaller in proportion to its total size and much cooler. In all likelihood it is not even molten, or only a portion of it is.

It appears that the Moon has an internal structure not dissimilar to that of Earth—that is, a crust, mantle, and core. The materials that make up the Moon, however, are quite different. Aluminum, calcium, iron, magnesium, titanium, potassium, and phosphorus have been found on the Moon, but it seems bereft of organic compounds, indicating that life never existed there. Moon rocks are primarily of basalt, or hardened lava, and breccia, soil, and rock fragments that have melted together.

The formation of these rocks must have occurred a long, long time ago, because the Moon has not had volcanic activity for several billion years. Thus, it is almost entirely "dead," lacking even a magnetic field of any significance. Moon rocks have only the faintest magnetic field, suggesting the absence of a significant molten core, which is what gives Earth its own magnetism. From the traces of magnetism found in these rocks, it is possible that the Moon had a magnetic field of some significance at one time, but that time is long past.

THE VULNERABLE MOON. Owing to the Moon's lack of sufficient gravitational force, it retains no atmosphere, and this means that it is completely vulnerable to particles reaching it from space. Early in its life, the Moon was subjected to a 700-million-year-long meteor shower that formed its craters. The damage from these showers was so great that it melted the Moon's crust, and eventually lava from the lunar interior surfaced to fill in cracks made by the meteorites.

In 1609 the Italian astronomer Galileo Galilei (1564–1642), the first scientist to gaze at the Moon through a telescope, observed dark patches that looked to him like bodies of water. He named these patches *seas,* a term that has remained in use among lunar cartographers, though Galileo's "seas" long ago were identified as dark spots made by cooling lava in the cracked surface.

As it is vulnerable to particles from space, the Moon also is susceptible to dramatic temperature changes brought about by the presence or absence of sunlight. For this reason, parts of the lunar surface are extremely hot at certain times, while other portions have never been penetrated by sunlight and are almost inconceivably cold.

HUMANS ON THE MOON

One such cold spot is at the Moon's south pole, in a basin carved out by an asteroid. There, temperatures are as low as –387°F (–233°C), within 40°C of absolute zero, or the temperature at which all molecular motion virtually ceases. By contrast, Pluto, which is so far from the Sun that the great star appears merely as a bright dot from there, has an average surface temperature that is warmer by 8°C. Yet in this inhospitable spot on the lunar surface, the U.S. probe *Lunar Prospector* in 1998 found something quite surprising: water.

Mixed in with dirt in the South Pole–Aitken Basin are ice crystals, which scientists speculate make up about 10% of the material in the surrounding area. Apparently moisture residue from comets that struck the Moon over the past three billion years, the ice offers intriguing possibilities. If a cost-effective method for extracting ice from the soil can be developed, colonization and exploration of the Moon may become a reality.

MOON LANDINGS. For now, however, human presence on the Moon is more a thing of the past than of the future. The first man-made spacecraft to visit the Moon was the Soviet *Luna 2* in 1959, and seven years later, *Luna 9* became the first such craft to land on the lunar surface. Its landing dispelled long-held fears that the Moon was awash in thick layers of dust; in fact, the lunar surface is a grayish soil, primarily composed of rock fragments, from 5 ft. to 20 ft. (1.5 m to 6 m) deep.

Though the Soviets were the first to reach the Moon in unmanned craft, the United States was the first (and so far the only) nation to put a man on the Moon. On July 20, 1969, in one of the most dramatic events of human history, the U.S. astronauts Neil Armstrong (1930–) and Edwin ("Buzz") Aldrin (1930–) walked on the lunar surface while Michael Collins (1930–) piloted *Apollo 11* in its orbit around the Moon.

Photographs from the Moon landing show the American flag extended, as though waving in a breeze as it would on Earth, but it actually was held in place mechanically, since there is no wind on the Moon. Aldrin and Armstrong demonstrated one of the most dramatic differences between the Moon and Earth: a decrease in weight. A man who weighs 200 lb. (90.72 kg) on Earth would weigh 33.04 lb. on the Moon and would therefore be much easier to lift. But he would be no easier to push from side to side, because his mass would not have changed.

Weight is the product of mass multiplied by the rate of gravitational acceleration and is therefore dependent on the gravitational force of the celestial body on which it is measured. Mass, on the other hand, does not vary anywhere in the universe. Thus, although 33.04 lb. is equal to 14.99 kg on Earth, the hypothetical astronaut described here would still have a mass of 90.72 kg.

Physicists had long known these facts about weight and mass, but footage of Armstrong and Aldrin bouncing around on the lunar surface provided a much more vivid demonstration. As it turned out, however, their foray to the Moon was the beginning of an all-too-brief chapter. Over the next three years, the United States conducted five more lunar landings as well as the failed 1970 mission designated *Apollo 13* (portrayed in the 1995 movie of that name), which never landed due to an onboard explosion. The last Moon mission took place in 1972, and soon afterward America was thrown into a recession spawned by the 1973 oil crisis. The next sustained effort at space flight, which began with the launch of the space shuttle in 1981, had an entirely different mission and destination—one that lay well within Earth's gravitational field.

SOLAR ENERGY, TEMPERATURE, AND COMPOSITION

Needless to say, there never will be any space missions, manned or otherwise, to the Sun. Spacecraft have passed close to Mercury, but as for the Sun, even its corona, or outermost atmospheric layer, has a temperature of 2,000,000 on the Kelvin scale of absolute temperature, or 3,599,541°F.

The Sun could not be more different from the Moon, most notably in terms of energy and temperature. It should be noted, however, that the Sun's temperature does not increase uniformly from the corona to the core. The "coolest part," at 5,800K (9,981°F), is the photosphere, a layer some 300 mi. (480 km) thick that constitutes the visible surface of the Sun. Above it, temperatures rise through the chromosphere, which is about 1,600 mi. (2,560 km) thick, becoming hotter still in the corona. Scientists do not understand the reasons for this temperature rise at the outer surface of the Sun.

Less surprising is the continued increase of temperature beneath the photosphere. The deeper inside the Sun, the higher the temperatures, until it reaches a staggering 15,000,000K (27,000,000°F) at the core. It is so hot there that not even atoms can exist; instead, the Sun's core is made up of subatomic particles—specifically, protons and electrons. Heat and movement are directly related in physics, and these particles are moving very fast—so fast that nuclear fusion can and does occur when these particles smash together.

NUCLEAR FUSION IN THE SOLAR CORE. When four protons fuse and absorb two electrons, amazing things happen. The result is the creation of a helium nucleus, whose mass is slightly smaller than the combined mass of the separate particles. Where did that "missing" mass go? It was converted to energy—an amount of energy that is staggering when multiplied by the large numbers of hydrogen atoms being converted to helium at any given moment.

That, in essence, is how the Sun creates energy—by converting hydrogen to helium. This energy is radiating from the Sun in electromagnetic waves at an amazing rate every second, and eventually it will be used up. Long before that happens, however, the Sun will begin to expand and cool. "Cooling," of course, is a relative term; this expanding Sun, a red giant, will burn up Earth even as it absorbs Mercury entirely. Then, when it has used up all its hydrogen, nothing will remain but a glowing core called a white dwarf.

There is no need to worry, however, because the events described will not happen anytime soon. The Sun has about as much life left in it as the amount of time it has lasted so far—approximately five billion years, or longer than Earth has existed. Given the fact that about 4.96 million tons (4.5×10^6 metric tons) of hydrogen are converted to helium every second, this gives some idea of the vast energy reserves on the Sun.

THE SUN'S ENERGY AND EARTH. Sunlight is more than just light, though that alone is a marvelous thing. The rays projected by the Sun are electromagnetic energy, of which visible light is only a small part. Also contained in the electromagnetic spectrum are long waves, short waves, and microwaves (including those used for transmitting radio and television signals); infrared and ultraviolet light; and x rays and gamma rays. Solar energy travels to Earth by means of radiation, a form of heat transfer that, unlike conduction or convection, requires no physical medium such as air. It can move through the vacuum of space to Earth's atmosphere, where a portion of it is absorbed and becomes the fuel that powers spaceship Earth.

It is a measure of the Sun's vast energy that Earth receives only 0.00000005% of its total output at any given moment. Of that small fraction (equal to one part in two billion), a much smaller portion makes it through Earth's atmosphere—yet that is enough to light the world and to facilitate the myriad other functions, such as photosynthesis, for which we depend on the Sun. Many effects of the Sun's light and energy are less than desirable, of course: sunburns and sometimes even tans, bleaching of materials exposed to light for too long, temporary blindness caused by gazing at the Sun or even at something reflecting it, and so on. These effects, too, attest to the Sun's awesome power.

The Sun makes possible the operations of three of four Earth systems and indirectly affects the fourth (see Earth Systems). The hydrosphere and atmosphere are both affected, for instance, by the evaporation of water for eventual precipitation, a process powered by the Sun. Likewise, the biosphere could not exist without photosynthesis and the other biological processes dependent on sunlight. Even the geosphere, though not directly powered by the Sun, is influenced by Sun-powered phenomena from the other three spheres.

CURIOUS SOLAR PHENOMENA

Even traveling at the speed of light, a photon takes nearly 30,000 years to travel from the center of the Sun to the corona. If it were in a vacuum, it could make the journey in about four seconds, but it continually bumps into other particles, and this slows it down (to put it mildly). Once it finally escapes the solar surface, it travels rather quickly, transmitting the Sun's light to the solar system.

When that light reaches Earth's atmosphere, it creates a number of strange optical effects. One of the best known is the rainbow, produced by the refraction, or bending, of light inside a raindrop. Because the colors of the visible spectrum are bent at different angles, they disperse as though in a prism; after being refracted a second time by the surrounding air, the billions of raindrops that fill the atmosphere after a storm produce a brilliant band of light.

When sunlight strikes oxygen or nitrogen, the two most significant elements in our atmosphere, the shorter wavelengths of light—green, blue, and violet—are the ones reflected. When

The aurora borealis, or northern lights. This phenomenon is caused by solar particles drawn to Earth's atmosphere, where they collide with oxygen and nitrogen molecules and become electrically charged, taking on a colored glow. *(© Michael Giannechini/Photo Researchers. Reproduced by permission.)*

these wavelenths combine, they appear bluish. The combination of light and differing temperatures on the ground produces a mirage, while ice crystals in the air may bring about haloes around the Sun or Moon and, in some cases, a sun dog, a reflected image of the Sun.

THE NORTHERN LIGHTS. One of the most breathtaking displays of atmospheric optical phenomena are auroras (or aurorae), natural fireworks that appear in the night sky. Visible in the northern United States and Canada, the auroras there are known as aurora borealis, or northern lights. A similar phenomenon occurs in the Southern Hemisphere, where it is called the southern lights, or aurora australis.

The source of the auroras is a stream of particles, called the *solar wind,* that emanates from the Sun. The solar wind is a consequence of the extreme heat on the Sun's surface, which causes atoms there to move so rapidly that even the enormous gravitational force of the Sun itself cannot hold them. Solar particles pass through the entire solar system and out to space beyond, but few of them reach Earth's atmosphere, because they are deflected by the planet's magnetic field.

The particles that cause auroras are electrons that instead of being deflected, are drawn toward Earth's polar regions. Two great rings of charged particles, called the Van Allen belts, are located high above each polar region, and they catch the majority of charged particles flowing toward them from the northern and southern hemispheres, respectively. The Van Allen belts cannot trap all the charged particles, however, and some enter the atmosphere, where they collide with oxygen and nitrogen molecules at about 50 mi. to 600 mi. (80 km to 1,000 km) above sea level. Some of these molecules become electrically charged and take on a colored glow. At higher levels oxygen glows red, while oxygen at lower levels is yellowish-green. Nitrogen glows blue.

SUNSPOTS AND THE SOLAR ACTIVITY CYCLE. Though auroras are visually remarkable, a more significant solar phenomenon is the sunspot. Sunspots are cooler regions—many of them vast in size—on the photosphere. They tend to be about 2,700°F (1,500°C) cooler than the surrounding areas, and for this reason they appear darker.

Sunspots seem to occur in 11-year cycles—a period known as a solar activity cycle—and are the result of strong magnetic fields. The latter, in turn, result from something called differential rotation: the Sun's equator rotates once every 26 days, whereas its poles rotate every 36 days, resulting in massive twisting of magnetic fields. This can produce anomalies such as sunspots, which disrupt radio communications on Earth.

Other solar anomalies, such as prominences, or hot spots formed by magnetic loops in the Sun's atmosphere, also follow the 11-year solar activity cycle. Actually, the cycle as it is known and measured *today* is about 11 years long. Its length seems to have varied over time, producing past changes in global temperature: the shorter the cycle, the warmer the temperatures on Earth.

The Sun, Moon, and Events on Earth

As noted earlier, the Sun and Moon have long inspired awe in humans, and in this the Moon has been a more than equal partner. Prehistoric and ancient humans regarded both lunar eclipses, in which Earth's shadow covers the Moon, and solar eclipses, in which the Moon comes between Earth and the Sun, as portents from the gods. From at least the era of the Babylonians onward, the Sun and Moon played complementary roles in marking time.

The marking of time according to the Sun relies on an objective reality, rather than a mere human construct. The year, which involves a complete cycle of seasons, is based on the amount of time Earth takes to revolve around the Sun. During this time, the planet also moves on its axis, causing the changes in orientation that bring about the seasons. Likewise a day is an objective reality, being the amount of time it takes Earth to revolve on its axis.

The month and week, on the other hand, relate to the phases of the Moon, which are themselves dependent on the Moon's position relative to Earth and the Sun. A lunar cycle lasts about 29.5 days, and this became the basis for the month, which lasts from 28 to 31 days. Within that lunar cycle are four phases—new, first quarter, full, and last quarter—which eventually became the basis for the idea that a month has about four seven-day weeks.

DAYS OF THE WEEK. Until about 1500, people thought of the Sun and Moon as two of the seven "planets" (including the five planets visible with the naked eye) that supposedly revolved around Earth. They further related these "planets" to the days of the week. As a result, virtually every culture that speaks a Euro-

KEY TERMS

ABSOLUTE ZERO: The temperature at which all molecular motion virtually ceases.

ASTRONOMICAL UNIT (AU): A figure equal to the average distance from Earth's center to the center of the Sun. The SI figure for an AU, adopted in 1996, is equal to $1.49597870691 \times 10^8$ km, or approximately 92,955,807 mi.

ATMOSPHERE: In general, an atmosphere is a blanket of gases surrounding a planet. Unless otherwise identified, however, the term refers to the atmosphere of Earth, which consists of nitrogen (78%), oxygen (21%), argon (0.93%), and other substances that include water vapor, carbon dioxide, ozone, and noble gases such as neon, which together comprise 0.07%.

ATOM: The smallest particle of an element, consisting of protons, neutrons, and electrons. An atom can exist either alone or in combination with other atoms in a molecule.

ATOMIC MASS UNIT: An SI unit (abbreviated amu), equal to 1.66×10^{-24} g, for measuring the mass of atoms.

ATOMIC NUMBER: The number of protons in the nucleus of an atom. Since this number is different for each element, elements are listed on the periodic table in order of atomic number.

AVERAGE ATOMIC MASS: A figure used by chemists to specify the mass—in atomic mass units—of the average atom in a large sample.

BIOSPHERE: A combination of all living things on Earth—plants, mammals, birds, reptiles, amphibians, aquatic life, insects, viruses, single-cell organisms, and so on—as well as all formerly living things that have not yet decomposed. Typically, after decomposing, a formerly living organism becomes part of the geosphere.

CORE: The center of Earth, which appears to be of molten iron. For terrestrial planets in general, *core* refers to the center, which in most cases is probably molten metal of some kind.

COSMOLOGY: The study of the origin, structure, and evolution of the universe.

COSMOS: The universe.

ELECTROMAGNETIC ENERGY: A form of energy with electric and magnetic components, which travels in waves.

ELECTROMAGNETIC SPECTRUM: The complete range of electromagnetic waves on a continuous distribution from a very low range of frequencies and energy levels, with a correspondingly long wavelength, to a very high range of frequencies and energy levels, with a correspondingly short wavelength. Included on the electromagnetic spectrum are long-wave and short-wave radio; microwaves; infrared, visible, and ultraviolet light; x rays, and gamma rays.

ELECTRON: A negatively charged particle in an atom, which spins around the nucleus.

ELEMENT: A substance made up of only one kind of atom. Unlike compounds, elements cannot be broken chemically into other substances.

GEOSPHERE: The upper part of Earth's continental crust, or that portion of the solid earth on which human beings live

KEY TERMS CONTINUED

and which provides them with most of their food and natural resources.

HYDROSPHERE: The entirety of Earth's water, excluding water vapor in the atmosphere but including all oceans, lakes, streams, groundwater, snow, and ice.

INERTIA: The tendency of an object in motion to remain in motion and of an object at rest to remain at rest.

ION: An atom that has lost or gained one or more electrons and thus has a net electric charge.

ISOTOPES: Atoms that have an equal number of protons, and hence are of the same element, but differ in their number of neutrons. This results in a difference of mass. An isotope may be either stable or radioactive.

JOVIAN PLANETS: The planets between Mars (the last terrestrial planet) and Pluto, all of which are large, low in density, and composed primarily of gases.

KELVIN SCALE: Established by William Thomson, Baron Kelvin (1824–1907), the Kelvin scale measures temperature in relation to absolute zero, or 0K. (Note that units in the Kelvin system, known as Kelvins, do not include the word or symbol for "degree.") The Kelvin scale, which is the system usually favored by scientists, is directly related to the Celsius scale; hence Celsius temperatures can be converted to Kelvin by adding 273.15.

LIGHT-YEAR: A unit of distance used by astronomers for measuring the extremely large expanses of space. Equal to the distance light travels in a year, a light-year is $9.460528405 \times 10^{12}$ km, or approximately 5.88 trillion mi.

LITHOSPHERE: The upper layer of Earth's interior, including the crust and the brittle portion at the top of the mantle.

MANTLE: The layer, approximately 1,429 mi. (2,300 km) thick, between Earth's crust and its core. In reference to the other terrestrial planets, *mantle* simply means the area of dense rock between the crust and core.

MASS: A measure of inertia, indicating the resistance of an object to a change in its motion. (By contrast, weight—which people tend to think of as analogous to mass—is a measure of gravitational force, or mass multiplied by the acceleration due to gravity.)

NEUTRON: A subatomic particle that has no electric charge. Neutrons are found at the nucleus of an atom, alongside protons.

NUCLEAR FUSION: A nuclear reaction that involves the joining of atomic nuclei.

pean language—not only in Europe itself but also among former European colonies in the Americas, Africa, Asia, and the Pacific—refers to the first day of the week as the Sun's day and the second as the Moon's day.

Today, in countries that speak Romance languages, or languages derived from Latin (most notably Italian, French, Spanish, and Portuguese), variations on the original Latin names remain in use: Domingo and Lunes in Spanish, for instance, or Dimanche and Lundi in French. Germanic languages adopted the ideas of the Latin day names, for the most part, but translated them into their own tongues: hence, Sonntag

KEY TERMS CONTINUED

NUCLEUS: The center of an atom, a region where protons and neutrons are located and around which electrons spin.

ORGANIC: At one time, chemists used the term "organic" only in reference to living things. Now the word is applied to most compounds containing carbon, with the exception of calcium carbonate (limestone) and oxides, such as carbon dioxide.

PERIODIC TABLE OF ELEMENTS: A chart that shows the elements arranged in order of atomic number, along with chemical symbol and the average atomic mass for that particular element.

PHOTON: A particle of electromagnetic radiation carrying a specific amount of energy.

PLANETARY SCIENCE: The branch of the earth sciences, sometimes known as planetology or planetary studies, that focuses on the study of other planetary bodies. This discipline, or set of disciplines, is concerned with the geologic, geophysical, and geochemical properties of other planets but also draws on aspects of astronomy, such as cosmology.

PROTON: A positively charged particle in an atom.

RADIATION: The transfer of energy by means of electromagnetic waves, which require no physical medium (for example, water or air) for the transfer. Earth receives the Sun's energy, via the electromagnetic spectrum, by means of radiation.

RADIOACTIVITY: A term describing a phenomenon whereby certain materials are subject to a form of decay brought about by the emission of high-energy particles or radiation. Forms of particles or energy include alpha particles (positively charged helium nuclei); beta particles (either electrons or subatomic particles called positrons); or gamma rays, which occupy the highest energy level in the electromagnetic spectrum.

REFRACTION: The bending of light as it passes at an angle from one transparent material into a second transparent material. Refraction accounts for the fact that objects under water appear to have a different size and location than they have in air.

SI: An abbreviation of the French term *Système International d'Unités,* or "International System of Units." Based on the metric system, SI is the system of measurement units in use by scientists worldwide.

TERRESTRIAL PLANETS: The four inner planets of the solar system: Mercury, Venus, Earth, and Mars. They are all small, rocky, dense, have relatively small amounts of gaseous elements, and are composed primarily of metals and silicates. Compare with Jovian planets.

and Montag in German, or Sunday and Monday in English.

CYCLES AND TIDES. One reason the ancients respected the Moon as much as the Sun is that it seemed to affect many aspects of human life—as indeed it does. Medieval ideas about "lunacy" were themselves more than a little off-kilter, and the belief that the Moon affects female menstrual cycles has come under significant challenge. Nevertheless, the Moon does seem to have some effect on human biological cycles.

The human circadian rhythm, or cycle of sleep and wakefulness, stretches over a period of

25 hours, meaning that a person living in a cave without any exposure to sunlight would eventually assume a 25-hour-a-day schedule. The reason for this disparity between the human circadian cycle and the length of a solar day is not known, but it is possible that the circadian cycle is based on the length of a lunar day, or the interval between periods of time when the Moon appears in the sky over a given spot on Earth. Because the Moon is moving even as Earth is rotating, this span of time is not 24 hours, the interval it takes Earth to rotate on its axis, but 24 hours and 50 minutes.

It is also more than an old wives' tale that people behave strangely around the time of a full moon: in fact, more deaths and accidents do occur at that point in the monthly cycle. One of the most significant lunar-influenced cycles, however, is not monthly but semidiurnal, or twice daily. These are the cycles of Earth's tides, brought about by the gravitational attraction exerted on Earth by the nearest significant celestial object. (The Sun also affects tides, but much less so because of its greater distance from Earth.) Though the Moon's effect on tides has been known since ancient times, the ancients did not understand the gravitational nature of its attraction, which actually causes a bulge to appear in the oceans.

In fact, two bulges appear, one in the oceans on the side of Earth nearest the Moon and one on the opposite side. The latter is also a result of the Moon's gravitation, which pulls Earth in the other direction, thus drawing the solid earth away from the water. These bulges are known as high tide, and the resulting displacement of water causes a low tide. In most places on Earth, tides are semidiurnal, meaning that there are two full cycles of high and low tide per day. Sometimes these variations can be very great, as at the Bay of Fundy in Canada, where the tidal range is as large as 46 ft. (14 m).

The tides are related closely to phases of the Moon, or, to put it another way, they are affected by the alignment between the Moon and the Sun. At the new moon, the Moon is between Earth and the Sun, and during a full moon, the Moon is on the other side of Earth from the Sun. (Because of a 20° angle between Earth's orbit and that of the Moon, Earth rarely casts a shadow on a full moon, except in the case of a lunar eclipse.) In the new-moon and full-moon phases, the Moon and Sun are aligned, and this combination of lunar and solar gravitation produces strong spring tides every 14 days. On the other hand, when the Moon and Sun are at angles to each other—during the first quarter and last quarter, sometimes known as half-moons—it produces a much weaker neap tide.

WHERE TO LEARN MORE

Beatty, J. Kelly, Carolyn Collins Petersen, and Andrew Chaikin. *The New Solar System.* 4th ed. New York: Sky Publishing, 1999.

Earth and Moon Viewer (Web site). <http://www.fourmilab.ch/earthview/vplanet.html>.

Lewis, John S. *Physics and Chemistry of the Solar System.* San Diego: Academic Press, 1997.

Lodders, Katharina, and Bruce Fegley. *The Planetary Scientist's Companion.* New York: Oxford University Press, 1998.

The Moon (Web site). <http://seds.lpl.arizona.edu/nineplanets/nineplanets/luna.html>.

Moon Phases (Web site). <http://www.astro.wisc.edu/~dolan/java/MoonPhase.html>

Moore, Patrick. *The Data Book of Astronomy.* Philadelphia: Institute of Physics, 2000.

Pérez, Miguel. *The Earth and the Universe: How the Sun, Moon, and Stars Cause Changes on Earth.* Illus. María Rius. Happage, NY: Barron's Educational Series, 1998.

Skinner, Brian J., Stephen C. Porter, and Daniel B. Botkin. *The Blue Planet: An Introduction to Earth System Science.* 2d ed. New York: John Wiley and Sons, 1999.

The Solar and Heliospheric Observatory (Web site). <http://sohowww.nascom.nasa.gov/>.

The Sun (Web site). <http://www.solarviews.com/eng/sun.htm>.

The Sun (Web site). <http://www.seds.org/nineplanets/nineplanets/sol.html>.

SCIENCE OF EVERYDAY THINGS

REAL-LIFE EARTH SCIENCE

GEOLOGY

HISTORICAL GEOLOGY

CONCEPT

Geologists are concerned primarily with two subjects: Earth's physical features and the study of the planet's history. These two principal branches of geology are known, appropriately enough, as physical geology and historical geology. Today they are of equal importance, but in the early modern era, geologists were most focused on topics related to historical geology, in particular, Earth's age and the means by which Earth was formed. This debate pitted adherents of religion, which seemed to require a very young Earth, against adherents of science. A breakthrough came with the introduction of uniformitarianism, a still-influential principle based on the idea that the geologic processes at work today have always been at work. Opposing uniformitarianism was catastrophism, or the idea that Earth was formed in a short time by a series of cataclysmic events. Discredited at the time, catastrophism later gained acceptance, though this did not lead to support for the concept of a young Earth. In fact, the planet is very old—so old that all of human history is almost inconceivably short in comparison.

HOW IT WORKS

EXPLAINING ORIGINS

For thousands of years, humans were content to rely on religiously inspired stories, rather than scientific research, to provide an explanation regarding Earth's origins. This topic, along with the scientific challenge to those early accounts of Earth's formation, is discussed in considerable detail within the essay Earth, Science, and Nonscience. Other aspects of the subject, particularly the challenge to mythological explanations put forward by earth scientists in modern times, are examined here.

All religious explanations of the planet's origins can be called myths, which is not necessarily a pejorative term: a myth is simply a story to explain how something came into being. So pervasive are myths about geology that a term, *geomythology,* has been coined to identify such myths. Geomythology in particular and mythology in general stand in sharp contrast to scientific explanations derived by using the scientific method of observation, hypothesis formation, testing of hypotheses, and the development and testing of theories.

CONTRASTING SCIENCE AND RELIGIOUS GEOMYTHOLOGY. Science and myth have in common the aim of explaining how things came to be, but the means by which they reach that explanation are quite different. So, too, are the reasons that drive science on the one hand and religion or myth on the other in seeking to develop such an explanation.

The most famous of all religious explanations of Earth's origins, of course, is that found in the biblical book of Genesis. Probably written in the latter part of the second millennium B.C., it offered a compelling story of creation that virtually defined the Western view of Earth's origins for more than 1,500 years, from about A.D. 300 to the beginning of the nineteenth century. Its purpose, of course, was not scientific, and it was not written as the result of research; rather, the Genesis account depicts nature as a vast stage on which a cosmic drama of love, sin, redemption, and salvation has been played out over the ages.

By contrast, the scientific search for Earth's origins is driven merely, or at least primarily, by curiosity to explain how things came to be as they are. Scientists certainly have their biases and are just as capable of error as anyone, yet at least they have a standard in the form of the scientific method. If a scientist's findings, and the resulting theory, withstand the rigorous testing required by the scientific method, the theory is rewarded with increasing acceptance and new research designed to test further its ability to explain the world. If the theory fails those tests, its adherents may hold on to it for a time, but eventually they die off, and the theory is discarded. On the other hand, adherence to the religious explanation of Earth's origins has proved more intractable, as we shall see.

The Religious War Concerning Earth's Origins

The great Italian artist and scientist Leonardo da Vinci (1452–1519) was among the first Western thinkers to speculate that fossils might have been made by the remains of long-dead animals. This was a daring supposition to make in the Renaissance, and it would become even more daring to uphold such an idea in the centuries that followed. The concept of fossils seemed to imply an Earth older than the biblical account suggested, and with the Catholic Church under attack by the forces of religious reformation (i.e., Protestantism) and other forms of "heresy," church leaders became less and less inclined to tolerate any deviation from orthodoxy.

The ecclesiastical view of Earth's history reached a sort of extreme in the seventeenth century, with the Irish bishop James Ussher (1581–1656). The New Testament contains a thorough accounting of Jesus' lineage, both through his mother and his earthly father, Joseph, all the way back to the time of Adam. Jesus was descended from David, whose lineage is provided in the Old Testament, complete with each ancestor's life span and the age at which he fathered a successor in the Davidic line of descent. From these figures, Ussher concluded that God finished making Earth at 9:00 A.M. on Sunday, October 23, 4004 B.C. Accepted by the Church, Ussher's calculation gave the idea of a very young Earth an aura of "scientific" justification.

Ironically, one of the first scientists to discover evidence that pointed toward an extremely old Earth was also a minister, the English astronomer Henry Gellibrand (1597–1636). While researching Earth's magnetic field, Gellibrand discovered that the field had changed over time (as indeed it has—for more on this subject, see Geomagnetism). This was one of the first indications that the planet's history can be studied scientifically, even though humans have no direct information regarding the origins of Earth.

Early Stratigraphic Studies

After Gellibrand came the Danish geologist Nicolaus Steno (1638–1687), who studied the age of rock beds. Thus was born the concept of stratigraphy, or the study of rock layers beneath Earth's surface, which revealed a great deal about the planet's age. (See Stratigraphy, which discusses many topics related to historical geology.) Along with the English physicist Robert Hooke (1635–1703) and others, Steno also became one of the first thinkers to confront the possibility that Earth must be much more than 6,000 years old.

During the eighteenth century, the German geologist Johann Gottlob Lehmann (1719–1767) built on ideas introduced by Steno concerning the formation of rock beds. Lehmann put forward the theory that certain groups of rocks tend to be associated with each other and that each layer of rock is a sort of chapter in the history of Earth. Aspects of Lehmann's theory were incorrect, but the general principle marked an advancement over previous ideas in geology and helped point the way toward a new view of the earth sciences.

Previously, geologic studies had tended to be qualitative and descriptive, meaning that earth scientists used very generalized terminology and failed to possess a grasp of larger issues. Thanks to Lehmann and others who followed him, the earth sciences became more truly quantitative and predictive, offering explanations of what had happened in the past, along with justifiable theories concerning what might happen in the future.

The German geologist Abraham Gottlob Werner (1750–1817) put forward a theory that was largely incorrect, yet one that nevertheless advanced the earth sciences. His "neptunist" theory was based on the idea that water had been the main force in shaping Earth's surface. Though this theory was not accurate, his idea was significant, because it constituted the first well-ordered geologic theory of Earth's origins and early history. At the same time, that history was turning out to be very long indeed.

LEONARDO DA VINCI WAS THE FIRST TO SUGGEST THAT FOSSILS ARE LONG-DEAD ANIMAL REMAINS, IMPLYING THAT EARTH IS OLDER THAN THE BIBLICAL ACCOUNT IN GENESIS WOULD SUGGEST. *(AP/Wide World Photos. Reproduced by permission.)*

RELIGION AND EARTH'S AGE

In 1774 the French mathematician Georges-Louis Leclerc, Comte de Buffon (1707–1788), applied new scientific ideas to the study of Earth and estimated its age at 75,000 years. Privately he admitted that he actually thought Earth was billions of years old but did not think that such a figure would be understood. At the time, after all, the concept of a "billion" was hardly a familiar one, as it is today. More important, the idea of

Earth being that old was shocking and downright frightening to people who accepted a strict interpretation of the biblical account.

In an earlier century, the Italian astronomer Galileo Galilei (1564–1642) had been forced, on pain of death, to recant his support for the Polish astronomer Nicolaus Copernicus's (1473–1543) discovery that Earth is not the center of the universe. In Buffon's day, by contrast, few Europeans faced such dire threats for endorsing apparently unbiblical ideas. A scientist could still lose his job for supporting the wrong principles, and thus Buffon had to renounce his position on threat of losing his post at the University of Paris.

AN ATHEISTIC REACTION. Other forces were at work in the sciences during the eighteenth century, and some were openly hostile to religious belief. An extreme example was the French physician and philosopher Julien de La Mettrie (1709–1751), a leading figure in the mechanist school of the biological sciences. La Mettrie maintained that humans are essentially a variety of monkey, to whom they were superior only by virtue of possessing the power of language. Moving far beyond the territory of science itself, he also taught that atheism is the only road to happiness and that the purpose of human life is to experience pleasure.

In the physical sciences, an interesting example of reaction to religious belief can be found in the case of the French mathematician Pierre Simon de Laplace (1749–1827). Like those of La Mettrie, Laplace's aims were not purely scientific; instead, he envisioned himself as a warrior against religious belief. Correctly enough, Laplace maintained that the origins of the universe as well as its workings could be explained fully without any reference to God. He also introduced a highly influential theory, widely accepted today, that the solar system originated from a cloud of gas. (See the entries Planetary Science and Sun, Moon, and Earth.)

Like La Mettrie, Laplace took his ideas far beyond their justifiable purview in the realm of science, however, wielding them as a sword in a religious war. Laplace maintained that because it was possible to discuss the origins of the cosmos without reference to God, there must be no God—which is far from a logically necessary conclusion. Misguided as La Mettrie's and Laplace's atheistic crusades may have been, they are historically understandable: in France, far more than anywhere in western Europe except perhaps Spain, the Church had come to be seen as a force of political oppression, allied as it was with the French royalty. It is no wonder, then, that the French Revolution of 1789 was directed as much against the Church as against the king.

REAL-LIFE APPLICATIONS

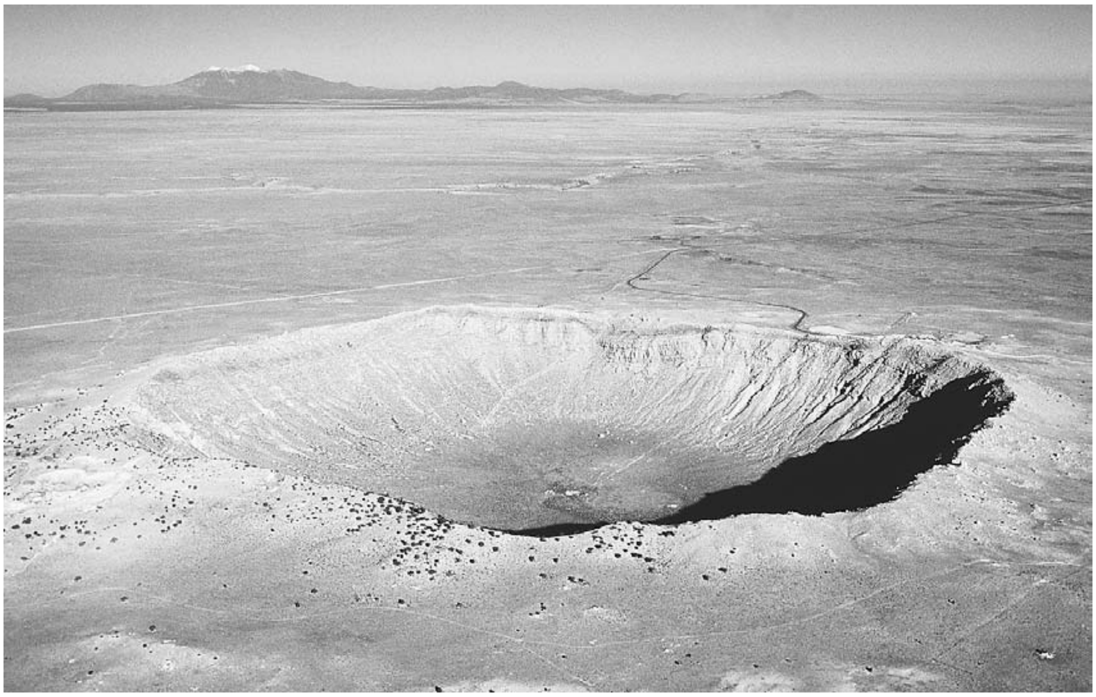
Uniformitarianism

Late in the eighteenth century, the Scottish geologist James Hutton (1726–1797) put forward an idea that transcended the debate over Earth's origins. Rather than speculate as to how Earth had come into being, Hutton analyzed the processes at work on the planet in his time and reasoned that they must be a key to understanding the means by which Earth was shaped. This was the principle of uniformitarianism, which is still a key concept in the study of Earth. Thanks to his introduction of this influential idea, Hutton today is regarded as the father of modern scientific geology.

Uniformitarianism, in general, is the idea that the geologic processes at work today provide a key to understanding the geologic past. This means that the laws of nature have always been the same. The uniformitarianism promoted by Hutton and his fellow Scottish geologist Charles Lyell (1797–1875), however, has undergone some modification, namely, by the addition of the qualifying statement that the speed and intensity of those processes may not always be the same at any juncture in geologic history. For instance, land does not erode today at the same rate that it did before plants existed to hold rocks and soil in place.

GOULD'S FOUR UNIFORMITIES. In the late twentieth century, the American paleontologist Stephen Jay Gould (1941–2002) identified four different meanings of uniformity in science, not all of which are equally valid. Gould's listing and analysis of these four meanings is as follows:

- Uniformity of law: The assumption that natural laws do not change over time. This idea governs all sciences.
- Uniformity of process: The idea embodied in the most well-known definition of geo-



METEOR CRATER, ARIZONA. MOST NOTABLE AMONG CATASTROPHE THEORIES OF EARTH'S FORMATION IS THE COLLISION OF METEORITES. *(© Francois Gohier/Photo Researchers. Reproduced by permission.)*

logic uniformitarianism, "The present is key to the past."

- Uniformity of rate: The incorrect assumption that the rate at which processes occur presently is the same as the rate at which they occurred in the past.
- Uniformity of state: The incorrect assumption that the state of the universe always has been as it is today.

As noted, uniformity of law is essential to all sciences. For instance, there is every reason to believe that the conservation of energy (a law stating that the total amount of energy in the universe remains constant) always has been the case. If the contrary were true, the conservation of energy could no longer properly be called a law, because it might cease to be the case at some time in the future.

The statement "The present is key to the past" was formulated by yet another Scottish geologist, Sir Archibald Geikie (1835–1924). Geikie's statement often has been criticized as an oversimplification, because processes that occurred in the past may not necessarily be occurring now, or vice versa, even though they could occur again. This idea has required modification of uniformitarianism, as noted earlier, to take into account the fact that the speed and intensity of processes may not always be the same. Part of this modification has involved acceptance of a form of catastrophism, discussed later in this essay.

Variations in the speed and intensity of processes also were addressed by Gould, with his observation that "uniformity of rate" is a fallacy. So, too, is "uniformity of state," which is one of the few areas on which adherents of creationism (a strict interpretation of the Genesis account) would agree with their opponents. Even the Bible, after all, says "In the beginning ... the earth was without form, and void."

CATASTROPHISM

In *Theory of the Earth* (1795), Hutton suggested that the weathering effects of water produced the sedimentary layers of Earth. Based on observation of river flow and mud content, he realized that this process would require much longer than 6,000 years. So, too, did Lyell, author of the highly influential *Principles of Geology,* which appeared in 12 editions from 1830 to 1875 and which presented a strict version of uniformitarianism.

Aqueducts and other structures erected by the Romans had stood for a good one-fourth to one-third of the entire history of Earth, assuming that it was as young as Ussher's biblical interpretation implied. Yet these Roman constructions had experienced very little weathering and certainly much less than mountains would have had to experience to leave behind the sediments observed by geologists. Surely, then, Earth must be millions upon millions of years old, not just a few thousand.

CUVIER'S CATASTROPHIC THEORY. Not so, countered adherents of a movement known as catastrophism, which arose in opposition to uniformitarianism during the late eighteenth and early nineteenth centuries. Catastrophism associates geologic phenomena with sudden, dramatic changes rather than ongoing and long-term processes, as in uniformitarianism. The leading proponent of catastrophism was the French geologist Baron Georges Cuvier (1769–1832), who used this theory to explain unconformities. These apparent gaps in the geologic record, revealed by observing rock layers, or strata, are discussed in the essay, Stratigraphy.

Whereas Cuvier's countryman (and fellow French noble) Buffon had asserted that Earth was 75,000 years old while actually believing that it was much older, Cuvier maintained that the planet is *just* 75,000 years old. The formation of mountains and other landforms, which should have taken millions of years, could be explained by sudden, violent changes, an example of which was Noah's Flood in the Book of Genesis. As the ocean waters receded, they moved rocks far from their sources, carved out valleys, and left behind lakes and other bodies of fresh water.

CATASTROPHISM TODAY. As more and more evidence for a very old Earth began to accumulate during the nineteenth century, catastrophism fell into disfavor. Discoveries from the 1970s onward, however, influenced a new look at catastrophism, and, as a result, the idea has received new attention in later years.

This has not led to a wholesale endorsement of creationism; rather, scientists have come to understand that the generally steady pace of processes on Earth periodically is broken by catastrophic events. Most notable among types of catastrophe is the collision of a meteorite with Earth, a remarkable example of which apparently occurred some 65 million years ago. That dramatic event seems to have forced so much dust and gas into the atmosphere that it blocked out the Sun, leading to the ultimate extinction of the dinosaurs.

Understanding Geologic Time

So just how old *is* Earth? Modern earth scientists working in the realm of historical geology, and specifically geochronology, estimate its age at about 4.6 billion years. (The dating techniques used to determine the age of the planet are discussed in the essay Stratigraphy.) Such a vast span of time is more than a little difficult for humans to comprehend, given the fact that our lives last 70–80 years, on average, and the entire history of human civilization is only about 5,500 years long.

For this reason, it is helpful to use scales of comparison, such as that offered at the Web site listed under the title *Comprehending Geologic Time.* Suppose that the entire geologic history of Earth were likened to a single year of 365.25 days, starting with the formation of the planet from a cloud of dust and ending with the present. More than two months would have been required simply for the accretion of Earth from a gas cloud to a planetesimal to something like its present form, but by about March 5 this evolution would have been accomplished.

The entire spring would be analogous to a long, long period of time in which Earth was pounded by meteor showers and the oceans began to form. Not even the oldest known rocks date back this far, and many of our ideas about this phase in Earth's history are based on conjecture. Much more is known about the second half of geologic history, beginning with the origins of the first single-cell life-forms on June 16.

FROM SINGLE CELLS TO DINOSAURS. We are now almost halfway through the year and still a long, long way from any sort of complex living beings. This is not surprising, given the fact that the formation of the continental plates and the development of oxygen in the atmosphere would have occurred only by about August 26. Even in the week after Thanksgiving, the most complex organisms would have been snails. Finally, a few days before the beginning of December, creatures would have begun to invade the land.

We tend to associate the dinosaurs with the early phases of Earth's history, but this only illustrates our distorted view of geologic time. In fact the Jurassic period, when dinosaurs roamed

KEY TERMS

CATASTROPHISM: The idea that geologic phenomena are brought about by sudden dramatic changes rather than ongoing and long-term processes, as in uniformitarianism. Although it was once used to promote the idea of a very young Earth, catastrophism today is accepted, in a very modified form, by many earth scientists.

GEOCHRONOLOGY: The study of Earth's age and the dating of specific formations in terms of geologic time.

GEOLOGY: The study of the solid earth, in particular, its rocks, minerals, fossils, and land formations.

GEOMYTHOLOGY: Folklore inspired by geologic phenomena.

HISTORICAL GEOLOGY: The study of Earth's physical history. Historical geology is one of two principal branches of geology, the other being physical geology.

QUALITATIVE: Involving a comparison between qualities that are not defined precisely, such as "fast" and "slow" or "warm" and "cold."

QUANTITATIVE: Involving a comparison between precise quantities—for instance, 10 lb. versus 100 lb. or 50 mi. per hour versus 120 mi. per hour.

SCIENTIFIC METHOD: A set of principles and procedures for systematic study that includes observation; the formation of hypotheses, theories, and laws; and continual testing and reexamination.

SEDIMENT: Material deposited at or near Earth's surface from a number of sources, most notably preexisting rock.

STRATIGRAPHY: The study of rock layers, or strata, beneath Earth's surface.

UNCONFORMITY: An apparent gap in the geologic record, as revealed by observing rock layers, or strata.

UNIFORMITARIANISM: The idea that the geologic processes at work today provide a key to understanding the geologic past. The speed and intensity of those processes, however, may not always be the same at any juncture in geologic history. Uniformitarianism usually is contrasted with catastrophism.

WEATHERING: The breakdown of rocks and minerals at or near the surface of Earth due to physical or chemical processes or both.

Earth, would be parallel to a period of about five days, from December 15 to 20. By Christmas Day, the meteorite referred to earlier would have hit Earth, and the dinosaurs would be headed toward extinction, their dead bodies eventually forming the fossil fuels that have powered much of human civilization.

THE SHORT SPAN OF HUMANITY'S EXISTENCE. By this point, we are within a few days of the year's end, and yet nothing remotely resembling a human has appeared. Our own species, *Homo sapiens,* would not have come on the scene until the last 0.16 days of the year—that is, at a few minutes after 8:00 p.m. on December 31. The New Year's Eve countdown would be nearing by the time human civilization began, at about 42 seconds before midnight.

Now we have come to a period about 6,000 years ago, or the point at which, according to Bishop Ussher, Earth was created. No wonder many people wanted to believe in a young Earth, and some even hold on to that belief today: when viewed against the backdrop of the planet's true age, humanity seems very insignificant indeed.

Christ's birth would have occurred at about 14 seconds before midnight, and the final 10-second countdown would begin about the time the Roman Empire fell. The life span of the average person would correspond to about half a second or less.

How Do We Know Earth's Age?

What we know about Earth's age comes, of course, not from direct observation but from the study of materials. One of the most important techniques for determining the age of samples taken from the earth is radiometric dating, discussed in more detail in Geologic Time. Radiometric dating involves ratios between two different kinds of atoms for a given element: stable and radioactive isotopes. Because chemists know how long it takes for half the isotopes in a given sample to stabilize (a half-life), they can judge the age of the sample by examining the ratio of stable to radioactive isotopes. In the case of uranium, one isotopic form, uranium-238, has a half-life of 4,470 million years, which is very close to the age of Earth itself. Use of uranium dating has detected rocks of an age between 3.8 and 3.9 billion years old, as well as even older crystal formations that suggest the earth had solid ground as early as 4.2 billion years ago.

A rock discovered in the Australian desert during the early 1980s appears to be the oldest rock sample in the world, according to data originally reported in *Nature* and included on the *Scientific American* Web site in early 2001. This zircon crystal, according to Simon Wilde of Curtin University in Western Australia, is 4.4 billion years old. Wilde and associates reported that extensive study of the sample suggested that at the time of its formation, Earth was already covered in water—something that had supposedly happened many millions of years later. If this was the case, it could suggest the possibility that life appeared much earlier than has previously been supposed, and perhaps even that life disappeared and reappeared several times before finally taking hold.

Where to Learn More

Bishop, A. C., A. Woolley, and A. Hamilton. *Cambridge Guide to Minerals, Rocks, and Fossils.* New York: Cambridge University Press, 1992.

Boggy's Links to Stratigraphy and Geochronology (Web site). <http://geologylinks.freeyellow.com/stratigraphy.html>.

Comprehending Geologic Time (Web site). <http://www.athro.com/geo/hgfr1.html>.

Harris, Nicholas, Alessandro Rabatti, and Andrea Ricciardi. *The Incredible Journey to the Beginning of Time.* New York: Peter Bedrick Books, 1998.

"Historical Geology." Georgia Perimeter College (Web site). <http://www.dc.peachnet.edu/~pgore/geology/geo102.htm>.

Lamb, Simon, and David Sington. *Earth Story: The Shaping of Our World.* Princeton, NJ: Princeton University Press, 1998.

MacRae, Andrew. *Radiometric Dating and the Geological Time Scale* (Web site). <http://www.talkorigins.org/faqs/dating.html>.

Reeves, Hubert. *Origins: Cosmos, Earth, and Mankind.* New York: Arcade, 1998.

Spickert, Diane Nelson, and Marianne D. Wallace. *Earthsteps: A Rock's Journey Through Time.* Golden, CO: Fulcrum Kids, 2000.

UCMP (University of California, Berkeley, Museum of Paleontology) Web Time Machine (Web site). <http://www.ucmp.berkeley.edu/help/timeform.html>.

GEOLOGIC TIME

CONCEPT

The expression geologic time refers to the vast span from Earth's beginnings to the present, about 4.6 billion years. To examine the history of Earth, one must discard most familiar ideas about time. Instead of thinking in terms of years, centuries, or even millennia, the most basic unit is a million years, and even that is rather small when compared with the four eons into which geologic time is divided. Earth scientists' knowledge of the first three eons is fairly limited. What they do know comes from a combination of absolute dating, mostly by the study of radioactive decay, and relative dating through the stratigraphic record of rock layers.

HOW IT WORKS

Historical Geology

The study of geologic time is encompassed within the larger subject of historical geology. The latter, the study of Earth's physical history, is one of the two principal branches of geology, the other being physical geology, or the study of Earth's physical components and the forces that have shaped them.

The background of historical geology is discussed in some detail within the Historical Geology essay. Its principal subdisciplines include stratigraphy, the study of rock layers, or strata, beneath Earth's surface; geochronology, the study of Earth's age and the dating of specific formations in terms of geologic time; sedimentology, the study and interpretation of sediments, including sedimentary processes and formations; paleontology, the study of fossilized plants and animals; and paleoecology, the study of the relationship between prehistoric plants and animals and their environments. Several of these subjects are examined in essays within this book.

Divisions of Geologic Time

Geologic time is divided according to two scales. The more well-known of these is the geologic scale, which divides time into named groupings according to six basic units: eon, era, period, epoch, age, and chron. In addition, the chronostratigraphic scale identifies successive layers of rock with specific units of time.

As noted earlier, stratigraphy is the study of rock layers, or strata, beneath Earth's surface, while chronostratigraphy is a subdiscipline devoted to studying the ages of rocks and what they reveal about geologic time. The chronostratigraphic scale likewise has six time units, analogous to those of the geologic scale: eonothem, erathem, system, series, stage, and chronozone. For the most part, we will not be concerned with the chronostratigraphic terms in the present context.

RELATIVE AND ABSOLUTE TIME. To discuss the divisions of geologic time, it is necessary first to discuss the concepts of relative and absolute time. The term *relative* refers to a quality or quantity that is comparative, or dependent on something else. Its opposite is *absolute,* a term designating a quality or quantity that is independent and not defined in relation to another quality or quantity.

If we say that Abraham Lincoln was born in 1809, it is an absolute designation of his birth year, whereas if we say that he was born 10 years after the death of George Washington (which

Subatomic particle tracks. Study of radioactive decay, or the release of subatomic particles, is a method of absolute dating. *(© John Giannicchi/Photo Researchers. Reproduced by permission.)*

occurred in 1799), that is an example of a relative time measurement. In actuality, of course, there is no truly absolute measure of time. For example, the reference to 1809 as Lincoln's birth year is based on the system of time measurement developed in the West, which, in turn, is based on early ideas regarding the date of Christ's birth. (As it turns out, Christ likely was born in about 6 B.C.)

Since the B.C./A.D. system of dating is widely accepted and used, or at least recognized, by most of the non-Western world, a date rendered according to this system constitutes the closest possible approximation to an absolute measure of time. In any case, one knows the difference between absolute and relative when one sees it: thus, to say that Lincoln was born ten years after Washington died is obviously and unmistakably a relative statement.

In terms of geology, the absolute age of a geologic phenomenon is its age in Earth years. On the other hand, its relative age is its age in comparison with other geologic phenomena, particularly the stratigraphic record of rock layers. Thus, references to relative age are given in terms of chronostratigraphic time divisions rather than millions of years.

Relative Dating

Given the meaning of relative age, it is easy enough to guess what relative dating would be, once one knows that dating, in a scientific context, usually refers to any effort directed toward finding the age of a particular item or phenomenon. Relative dating, then, assigns an age relative to that of other items, whereas absolute dating determines the age in actual years or millions of years.

One of the principal means of relative dating is through stratigraphy, which is based on the assumption that the deeper a layer of rock lies beneath Earth's surface, the earlier it was deposited. This holds true, however, for only one of the three major types of rock: sedimentary rock, which is formed by compression and deposition (i.e., formation of deposits) on the part of rock and mineral particles. (The other types of rock are igneous and metamorphic.)

Aside from stratigraphy, discussed in a separate essay, other relative dating techniques include seriation, faunal dating, and pollen dating, or palynology. Used, for instance, in archaeological studies, seriation analyzes the abundance of a particular item (for instance, pieces of pottery) and assigns relative dates based on this abundance. The term faunal dating refers to fauna, or animal life, and faunal dating is the use of animal bones to determine age. Finally, pollen dating, or palynology, involves analysis of pollen deposits.

Absolute Dating

As dating technology has progressed, it has become increasingly possible for scientists to provide absolute dates for specimens. One such method, introduced in the 1960s, is amino-acid racimization. Amino acids exist in two forms, designated *L*-forms and *D*-forms, which are stereoisomers, or mirror images of each other. Virtually all living organisms (except some microbes) incorporate only the L-forms, but once the organism dies the L-amino acids gradually convert to D-amino acids. Several factors influence the rate of conversion, and though amino-acid racimization was popular in the 1970s, these uncertainties have led scientists to treat it with increasing disfavor.

The principles that undergird amino-acid racimization, however, are essential to most forms of absolute dating. Generally, absolute dating uses ratios between the quantities of a particular substance (let us call it *Substance A*) and the quantities of a mirror substance (*Substance B*) to which it is converted over a period of time. The greater the ratio of Substance B to Substance A, the longer the time that has elapsed. The scale of time for various substances, however, differs greatly. Carbon-14 decay, for instance, takes place over a few thousand years, making it useful for measuring the age of human artifacts. On the other hand, uranium decay takes billions of years, and thus it is used for dating rocks.

Cation-ratio dating, for instance, measures the amount of cations, or positively charged ions, that have formed on an exposed rock surface. (An ion is an atom or group of atoms that have lost or gained electrons, thus acquiring a net electric charge. Electron loss creates a cation, as opposed to a negatively charged anion, created when an atom or atoms gain electrons.) Cation-ratio dating is based on the idea that the ratio of potassium and calcium cations to titanium cations decreases with age. It is applicable only to rocks in desert areas, where the dry air stabilizes the cation "varnish."

RADIOACTIVE DECAY. Various forms of radiometric dating employ ratios as well. Every element has a particular number of protons, or positively charged particles, in its nucleus, but it may have varying numbers of neutrons, particles with a neutral electric charge but relatively great mass. (Neutrons and protons have approximately the same mass, which is more than 1,800 times greater than that of an electron.) When two or more atoms of the same element have a differing number of neutrons, they are called isotopes.

Some types of isotopes "fit" better with a particular element and tend to be most abundant. For instance, carbon has six protons, and it so happens that the most abundant carbon isotope has six neutrons. Because there are six protons and six neutrons, totaling 12, this carbon isotope is designated *carbon-12,* which accounts for 98.9% of the carbon in nature. Generally speaking, the most abundant isotope is also the most stable one, or the one least likely to release particles and thus change into something else.

This release of particles is known as radioactive decay. In the context of radioactivity, "to decay" does not mean "to rot" rather, the isotope expels alpha particles (positively charged helium nuclei), beta particles (either electrons or subatomic particles called positrons), or gamma rays, which occupy the highest energy level in the electromagnetic spectrum. In so doing, it eventually will become another isotope, either of the same element or of a different element, and will stabilize. The amount of time it takes for half the isotopes in a sample to stabilize is called its half-life. This half-life varies greatly between isotopes, some of which have a half-life that runs into the billions of years.

DETERMINING ABSOLUTE AGE

When an organism is alive, it incorporates a certain ratio of carbon-12 in proportion to the amount of the radioisotope (that is, radioactive isotope) carbon-14 that it receives from the atmosphere. As soon as the organism dies, however, it stops incorporating new carbon, and the ratio between carbon-12 and carbon-14 will begin to change as the carbon-14 decays to form nitrogen-14. A scientist can use the ratios of carbon-12, carbon-14, and nitrogen-14 to ascertain the age of an organic sample.

Carbon-14, known as radiocarbon, has a half-life of 5,730 years, meaning that it takes that long for half the isotopes in a sample to decay to nitrogen-14. Note that half-life is *not* half the amount of time it takes for the entire sample to decay, especially because the first half of the sample usually decays faster than the second half. Imagine, for instance, that you had 100 units and wanted to reduce it to zero units by continually halving it. At first, the results would be dramatic, as 100 became 50, then 25, then 12.5, and so on. Eventually you would be down to smaller and smaller fractions of 1, and each division by 2 would yield a smaller number—but never zero.

Radioactive decay works that way as well, and, thus, while carbon-14 has a half-life of less than 6,000 years, it takes much longer than 6,000 years for the other half of the isotopes in a carbon-14 sample to decay. For this reason, the use of proper instrumentation makes it possible to judge the age of charcoal, wood, and other biological materials over a span of as long as 70,000 years. While this may be useful for archaeologists, it is not very helpful for measuring the vast spans of time encompassed in the earth sciences. Furthermore, there is a good likelihood that the sample will become contaminated by additional carbon from the soil. Moreover, it cannot be said with certainty that the ratio of carbon-12 to carbon-14 in the atmosphere has been constant throughout time.

POTASSIUM-ARGON DATING. Much more useful, from the standpoint of geology, is potassium-argon dating. When volcanic rocks are subjected to extremely high temperatures, they release the element argon, a noble gas. As the rocks cool, the stable isotope argon-40 accumulates. Because argon-40 is formed by the radioactive decay of a potassium isotope, potassium-40, the amount of argon-40 that forms is proportional to the rate of decay for potassium-40.

Potassium-40 has a half-life of 1.3 billion years, and with the help of argon-40, geologists have been able to estimate the age of volcanic layers above and below fossil and artifact remains in eastern Africa. Potassium-argon dating is most effective for rocks that are at least three million years old, because it takes about that long to accumulate enough argon-40 to make accurate measurements possible.

This brings up a notable aspect of radiometric dating techniques. No one technique is most effective; rather, each technique is suited to a particular span of time. Thus, potassium-argon dating would be virtually useless for measuring the relatively short time scales for which radiocarbon dating is ideally suited. The converse is also true: as we have noted, radiocarbon dating simply does not cover a wide enough span of time to be useful in most geologic studies.

URANIUM-SERIES DATING. We now come to the element most useful for dating the age of material samples over a broad chronological spectrum: uranium, which has an atomic number of 92. This means that it has 92 protons in its nucleus, making uranium atoms typically the heaviest atoms that occur in nature. (There are about 20 elements with atomic numbers higher than 92, but all of them have been created artificially, either in laboratories or as the result of nuclear testing.)

Both uranium and thorium, with an atomic number of 90, have unstable "parent" isotopes that decay into even more unstable "daughter" isotopes before eventually stabilizing as isotopes of lead. These daughter isotopes have half-lives that range from just a few years to a few hundred thousand years, whereas the half-lives of the parent isotopes are much longer. That of uranium-235, for instance, is 7.038×10^8 years, or more than 700 million years. On the other hand, the daughter isotope protactinium-231 has a half-life of 32,760 years.

When uranium-235 is deposited in an area, over time it will decay to form daughter isotopes. Assuming that the sample has been left undisturbed (isotopes have neither entered nor exited the deposit since its initial formation), the age of certain types of sample may thus be determined. For mollusks and corals, for instance, the amount of protactinium-231, a daughter isotope that begins to accumulate only after the organism dies, makes it possible to date a sample. In some cases, large amounts of a daughter isotope may be deposited initially alongside samples of a parent, and if these are present in water, the quantities of each can be judged according to the amount that has dissolved. For example, the daughter isotope uranium-234 dissolves more readily in water than the parent, uranium-238.

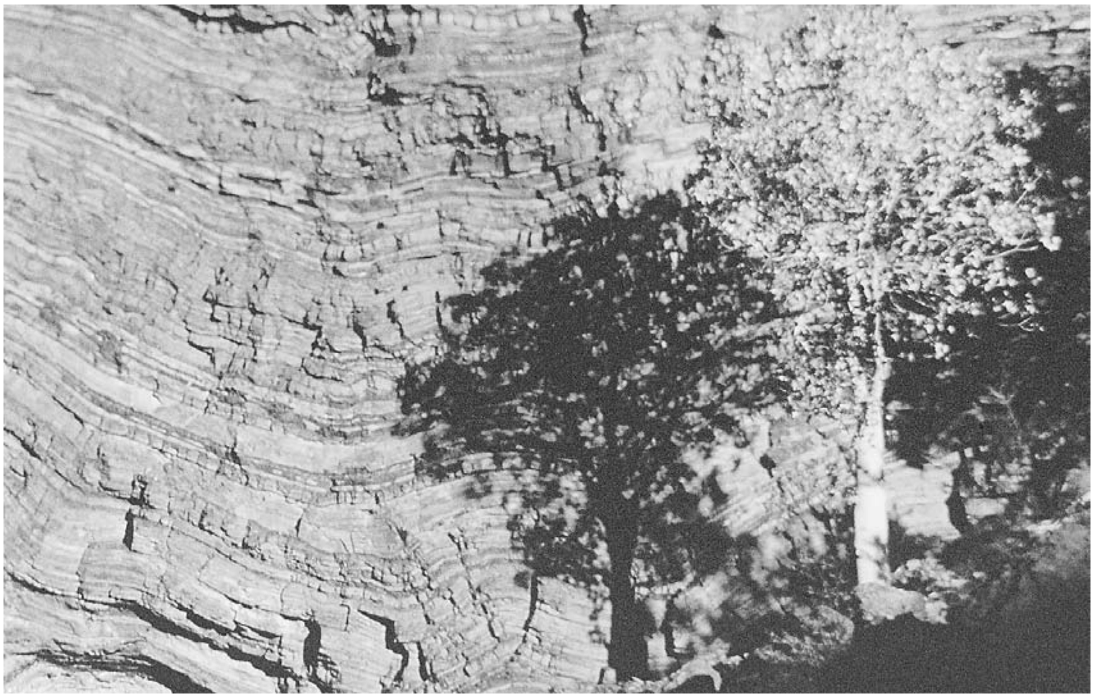
REAL-LIFE APPLICATIONS

How Do We Know Earth's Age?

We can now begin to answer a question almost inevitably raised when discussing geologic time: just how do we know that Earth is about 4.6 billion years old? A clue lies in the half-life of uranium-238, which is 4.47×10^9 years, or 4,470 million years. Geologists typically would abbreviate this as 4.47 Ga, the latter referring to "gigayears," a unit of a billion years.

As uranium atoms undergo fission, or splitting, this process releases energy that causes marks, called tracks, to form on the surface of volcanic minerals. In splitting, two daughter atoms shoot away from each other, forming tracks, and thus the rate of track formation is proportional to the rate of decay on the part of the parent isotope.

Incidentally, fission-track dating with uranium-238 defies the statement made earlier that certain types of dating are more suited to long periods of time, while others are best for shorter periods. When heated, the tracks disappear from



Relative dating uses the stratigraphic record of rock layers, as seen in the wall of this gorge, to determine the age of a geologic formation compared with others. *(© B. Bachman/Photo Researchers. Reproduced by permission.)*

a sample containing uranium-238, thus resetting the dating clock. As a result, if an object was heated just a few decades ago, it can be dated; so, too, can meteorites billions of years old.

Most meteorites found in the solar system tend to be about 4.56 Ga; hence, the rough figure of 4.6 Ga is used for the point at which the solar system, including Earth, began to form. The oldest known materials on Earth are zircon crystals from western Australia, dated about 4.3 Ga. Small samples of gneiss in Canada's Northwest Territories have been dated to about 4.0 Ga, but the oldest large-scale sample is a belt of 3.8 Ga gneiss in western Greenland.

A Question of Scale

Having discussed at least the rudiments of the dating system used by geologists, it is possible to examine geologic time itself. This requires a mental adjustment of monumental proportions, because one must discard all notions used in studying the history of human civilization. Concepts such as medieval, ancient, and prehistoric are practically useless when discussing geologic time, which dwarfs the scale of human events.

Human civilization has existed for about 5,500 years, the blink of an eye in geologic terms. Even the span of time that the human species, *Homo sapiens,* has existed—about two million years—is negligible in the grand scheme of Earth's history. The latter stretches back some 4,600 million years, meaning that human beings have existed on this Earth for just 0.043% of the planet's history. As discussed in the essay Historical Geology, if the entire history of Earth were likened to a single year, humans would have appeared on the scene at a few minutes after 8:00 p.m. on December 31. Human civilization would date only from about 42 seconds before midnight, and the age of machinery and industrialization would not fill up even the final two seconds of the year.

ANOTHER ANALOGY: LOS ANGELES TO NEW YORK. When discussing distances in space, astronomers dispense with miles, because they would be useless, given the vastness of the scale involved. The same is true of geologic time, in which the concept of years is hardly relevant. Instead, geologists speak in terms of millions of years, or megayears, abbreviated "Ma." (Geologists also use the much larger unit of a gigayear, to which we have already referred.) To discuss the age of Earth in terms of years, in fact, would be rather like measuring the distance from Los Angeles to New York in feet; instead, of course, we use miles. Now let us consider geologic time in terms of the 2,462 mi. between Los Angeles and New York, with 1 mi. equal to 1.8684 Ma, or 1,868,400 years.

Suppose we have left Los Angeles and driven a good deal of the distance to New York—46% of the way, in fact, to western Nebraska, a spot analogous to the beginning of the Proterozoic era. In the preceding miles, a duration equivalent to about 1,133 Ma, Earth was formed from a cloud of gas, pounded by meteors, and gradually became the home to oceans—but no atmosphere resembling the one we know now. The end of the Proterozoic era (about 545 Ma, or 545 million years ago) would be at about 88% of the distance from Los Angeles to New York—somewhere around Pittsburgh, Pennsylvania. By this point, the continental plates have been formed, oxygen has entered the atmosphere, and soft-bodied organisms have appeared.

We are a long way from Los Angeles, and yet almost the entire history of life on Earth, at least in terms of relatively complex organisms, lies ahead of us. If we skip ahead by about 339 Ma (a huge leap in terms of biological development), we come to the time when the dinosaurs appeared. We are now 95% of the way from the beginning of Earth's history to the present, and if measured against the distance from Los Angeles to New York, this would put us at a longitude equivalent to that of Baltimore, Maryland. Another 89 mi. would put us at about 65 million years ago, or the point when the dinosaurs became extinct.

We would then have only 33.7 mi. to drive to reach the point where humans appeared, by which time we would be in the middle of Manhattan. Compared with the distance from Los Angeles to New York, the span of human existence would be much smaller than the cab ride from Central Park to the Empire State Building. The entire sweep of written human history, from about a thousand years before the building of the pyramids to the beginning of the third millennium A.D., would be much smaller than a city block. In fact, it would be about the width of a modest storefront, or 15.54 ft.

The Very, Very Distant Past

So what happened for all those hundreds of millions of years before humans appeared on the scene? We will attempt to answer that question in an extremely cursory, abbreviated fashion, but for further clarification, the reader is strongly encouraged to consult a chart of geologic time. Such a chart can be found in virtually any earth sciences textbook; indeed, several versions (including a chronostratigraphic chart) may appear in a single book.

In addition to showing geologic time in both absolute and relative terms, these charts typically provide information about the magnetic polarity over a given span, since that has changed many times since Earth came into existence. In other words, what is today the magnetic North Pole was once the magnetic South Pole, and vice versa. (For more on this subject, see the discussion of paleomagnetism in the entries Plate Tectonics and Geomagnetism.)

As one might expect, disagreement between earth scientists is greatest with regard to the most distant phases of Earth's geologic history. This encompasses nearly 90% of all geologic time, dating back to about 545 Ma, thus showing how little geologists know, even today, about the geologic events of the very distant past. For this reason, when discussing Precambrian time, it is usually necessary to consider only the three eons that composed it. Discussion of era and period, on the other hand, is reserved for the three eras, and 11 periods, of the Phanerozoic eon. The smaller division of epoch is generally only of concern with regard to the most recent era, the Cenozoic. As for divisions smaller than an epoch, these will not concern us here.

THE PRECAMBRIAN EONS. The last paragraph of the preceding section encompasses a number of ideas, which now need to be explained, in at least general terms. The term *Precambrian* encompasses about four billion years of Earth's history, including three of the four eons (Hadean or Priscoan, Archaean, and Proterozoic) of the planet's existence. The names of these eons are derived from Greek, with the first being taken from the name of the deity who ruled over the Underworld. The latter two are derived, respectively, from the Greek words for *beginning* and *new life.*

The Hadean eon (sometimes called the Priscoan) lasted from about 4,560 Ma to 4,000

Ma ago, when the planet was being formed, or accreted, as pieces of solid matter floating around in the young solar system began to join one another. Meteorites showered the planet, bringing both solid matter and water, and thus forming the basis of the oceans. There was no atmosphere as such, but by the end of the eon, volcanic activity had ejected enough carbon dioxide and other substances into the air to form the beginnings of one. The oceans began to cool, making possible the beginnings of life—that is, molecules of carbon-based matter that were capable of replicating themselves. These appeared at the end of the Hadean eon, perhaps arriving from space in a meteorite.

The boundaries of the Precambrian eons are far from certain, so it is possible only to say that the Archaean eon lasted from about 4,000 Ma to 2,500 Ma ago. The earliest known datable materials, described earlier, all come from this time; in fact, outcrops of Archaean rock have been found on all seven continents. The rocks of this eon contain the first clear evidence of life, in the form of microorganisms. Over the course of the Archaean eon, prokaryotes, or cells without a nucleus, made their appearance, and later they were followed by eukaryotes, or cells with a nucleus.

During this great span of time, more than 20% of Earth's history, the atmosphere and hydrosphere developed considerably, even as the biosphere had its true beginnings. As for the geosphere, it also matured enormously in the course of the Archaean eon. During the Hadean eon, Earth's interior had begun to differentiate into core, mantle, and crust, and cooling in the two upper layers influenced the beginnings of the earliest plate-tectonic activity (see Plate Tectonics).

Even longer was the Proterozoic eon, which appears to have lasted from about 2,500 Ma to 545 Ma. This phase saw the beginnings of very basic forms of plant life, such that photosynthesis (the biological conversion of electromagnetic energy from the Sun into chemical energy in plants) began to take place. Plate-tectonic processes accelerated as well, with continents moving about over Earth's surface and smashing against one another. Oxygen in the atmosphere assumed about 4% of its present levels, but animal life still consisted primarily of eukaryotes.

THE PHANEROZOIC ERAS AND PERIODS. The end of the Proterozoic eon, once again, is not sharply defined in the stratigraphic record, such that there is considerable dispute as to the time periods involved. In any case, it is clear that the pace of development in the biosphere increased dramatically in the Phanerozoic, the eon in which we are now living. During the beginning of the Phanerozoic eon, algae appeared, and there followed an acceleration in the development of living organisms that ultimately produced the varied biosphere we know today.

As noted earlier, the only eras and periods that need concern most students of the earth sciences are those of the Phanerozoic eon. The three eras are as follows:

Eras of the Phanerozoic Eon

- Paleozoic (about 545 to 248.2 Ma)
- Mesozoic (about 248.2 to 65 Ma)
- Cenozoic (about 65 Ma to the present)

Within these eras are the following periods:

Periods of the Paleozoic Era

- Cambrian (about 545 to 495 Ma)
- Ordovician (about 495 to 443 Ma)
- Silurian (about 443 to 417 Ma)
- Devonian (about 417 to 354 Ma)
- Carboniferous (about 354 to 290 Ma)
- Permian (about 290 to 248.2 Ma)

Periods of the Mesozoic Era

- Triassic (about 248.2 to 205.7 Ma)
- Jurassic (about 205.7 to 142 Ma)
- Cretaceous (about 142 to 65 Ma)

Periods of the Cenozoic Era

- Palaeogene (about 65 to 23.8 Ma)
- Neogene (about 23.8 to 1.8 Ma)
- Quaternary (about 1.8 Ma to present)

These divisions, as well as the two most recent epochs of the Quaternary period (Pleistocene and Holocene), are discussed elsewhere in this book. It should be noted that there are variations for many of the eon, era, and period names given here; also, the Palaeogene and Neogene are often grouped together as a subera called the Tertiary. The latter nomenclature fits with a mnemonic device used by geology students memorizing the names of the 11 Phanerozoic periods: "Camels Ordinarily Sit Down Carefully; Perhaps Their Joints Creak Tremendously Quietly."

KEY TERMS

ABSOLUTE AGE: The absolute age of a geologic phenomenon is its age in Earth years. Compare with *relative age.*

ATMOSPHERE: In general, an atmosphere is a blanket of gases surrounding a planet. Unless otherwise identified, however, the term refers to the atmosphere of Earth, which consists of nitrogen (78%), oxygen (21%), argon (0.93%), and other substances that include water vapor, carbon dioxide, ozone, and noble gases such as neon, which together comprise 0.07%.

ATOM: The smallest particle of an element, consisting of protons, neutrons, and electrons. An atom can exist either alone or in combination with other atoms in a molecule.

BIOSPHERE: A combination of all living things on Earth—plants, animals, birds, marine life, insects, viruses, single-cell organisms, and so on—as well as all formerly living things that have not yet decomposed.

CHRONOSTRATIGRAPHY: A subdiscipline of stratigraphy devoted to studying the ages of rocks and what they reveal about geologic time.

DATING: Any effort directed toward finding the age of a particular item or phenomenon. Methods of geologic dating are either relative (i.e., comparative and usually based on rock strata) or absolute. The latter, based on such methods as the study of radioactive isotopes, usually is given in terms of actual years or millions of years.

ELECTRON: A negatively charged particle in an atom, which spins around the nucleus.

ELEMENT: A substance made up of only one kind of atom. Unlike compounds, elements cannot be chemically broken into other substances.

EON: The longest phase of geologic time. Earth's history has consisted of four eons, the Hadean or Priscoan, Archaean, Proterozoic, and Phanerozoic. The next-smallest subdivision of geologic time is the era.

EPOCH: The fourth-longest phase of geologic time, shorter than an era and longer than an age and a chron. The current epoch is the Holocene, which began about 0.01 Ma (10,000 years) ago.

ERA: The second-longest phase of geologic time, after an eon. The current eon, the Phanerozoic, has had three eras, the Paleozoic, Mesozoic, and Cenozoic, which is the current era. The next-smallest subdivision of geologic time is the period.

GA: An abbreviation meaning "gigayears," or "billion years." The age of Earth is about 4.6 Ga.

GEOCHRONOLOGY: The study of Earth's age and the dating of specific formations in terms of geologic time.

GEOLOGIC TIME: The vast stretch of time over which Earth's geologic development has occurred. This span (about 4.6 billion years) dwarfs the history of human existence, which is only about two million years. Much smaller still is the span of human civilization, only about 5,500 years.

GEOSPHERE: The upper part of Earth's continental crust, or that portion of the solid earth on which human beings live

KEY TERMS CONTINUED

and which provides them with most of their food and natural resources.

HISTORICAL GEOLOGY: The study of Earth's physical history. Historical geology is one of two principal branches of geology, the other being physical geology.

HYDROSPHERE: The entirety of Earth's water, excluding water vapor in the atmosphere but including all oceans, lakes, streams, groundwater, snow, and ice.

ISOTOPES: Atoms that have an equal number of protons and hence are of the same element but differ in their number of neutrons. This results in a difference of mass. An isotope may be either stable or radioactive.

MA: An abbreviation used by earth scientists, meaning "million years" or "megayears." When an event is dated to, for instance, 160 Ma, it usually means that it took place 160 million years ago.

NEUTRON: A subatomic particle that has no electric charge. Neutrons are found at the nucleus of an atom, alongside protons.

NUCLEUS: The center of an atom, a region where protons and neutrons are located and around which electrons spin.

PERIOD: The third-longest phase of geologic time, after an era. The current eon, the Phanerozoic, has had 11 periods, and the current era, the Cenozoic, has consisted of three periods, of which the most recent is the Quaternary. The next-smallest subdivision of geologic time is the epoch.

PRECAMBRIAN TIME: A term that refers to the first three of four eons in Earth's history, which lasted from about 4,Ma to about 545 Ma ago.

PROTON: A positively charged particle in an atom.

RADIOACTIVITY: A term describing a phenomenon whereby certain materials are subject to a form of decay brought about by the emission of high-energy particles or radiation. Forms of particles or energy include alpha particles (positively charged helium nuclei), beta particles (either electrons or subatomic particles called *positrons*, or gamma rays, which occupy the highest energy level in the electromagnetic spectrum.

RADIOMETRIC DATING: A method of absolute dating using ratios between "parent" isotopes and "daughter" isotopes, which are formed by the radioactive decay of parent isotopes.

RELATIVE AGE: The relative age of a geologic phenomenon is its age in comparison with other geologic phenomena, particularly the stratigraphic record of rock layers. Compare with *absolute age.*

SEDIMENT: Material deposited at or near Earth's surface from a number of sources, most notably preexisting rock.

SEDIMENTARY ROCK: Rock formed by compression and deposition (i.e., formation of deposits) on the part of other rock and mineral particles.

SEDIMENTOLOGY: The study and interpretation of sediments, including sedimentary processes and formations.

STRATIGRAPHY: The study of rock layers, or strata, beneath Earth's surface.

WHERE TO LEARN MORE

Boggy's Links to Stratigraphy and Geochronology (Web site). <http://geologylinks.freeyellow.com/stratigraphy.html>.

Comprehending Geologic Time (Web site). <http://www.athro.com/geo/hgfr1.html>.

Hancock, Paul L., and Brian J. Skinner. *The Oxford Companion to the Earth.* New York: Oxford University Press, 2000.

Harris, Nicholas, Alessandro Rabatti, and Andrea Ricciardi. *The Incredible Journey to the Beginning of Time.* New York: Peter Bedrick Books, 1998.

"Historical Geology." Georgia Perimeter College (Web site). <http://www.dc.peachnet.edu/~pgore/geology/geo102.htm>.

Lamb, Simon, and David Sington. *Earth Story: The Shaping of Our World.* Princeton, NJ: Princeton University Press, 1998.

MacRae, Andrew. *Radiometric Dating and the Geological Time Scale* (Web site). <http://www.talkorigins.org/faqs/dating.html>.

Reeves, Hubert. *Origins: Cosmos, Earth, and Mankind.* New York: Arcade, 1998.

Spickert, Diane Nelson, and Marianne D. Wallace. *Earthsteps: A Rock's Journey Through Time.* Golden, Colo.: Fulcrum Kids, 2000.

UCMP (University of California, Berkeley, Museum of Paleontology) Web Time Machine (Web site). <http://www.ucmp.berkeley.edu/help/timeform.html>.

STRATIGRAPHY

CONCEPT

Stratigraphy is the study of rock layers (strata) deposited in the earth. It is one of the most challenging of geologic subdisciplines, comparable to an exacting form of detective work, yet it is also one of the most important branches of study in the geologic sciences. Earth's history, quite literally, is written on the strata of its rocks, and from observing these layers, geologists have been able to form an idea of the various phases in that long history. Naturally, information is more readily discernible about the more recent phases, though even in studying these phases, it is possible to be misled by gaps in the rock record, known as unconformities.

HOW IT WORKS

The Foundations of Stratigraphy

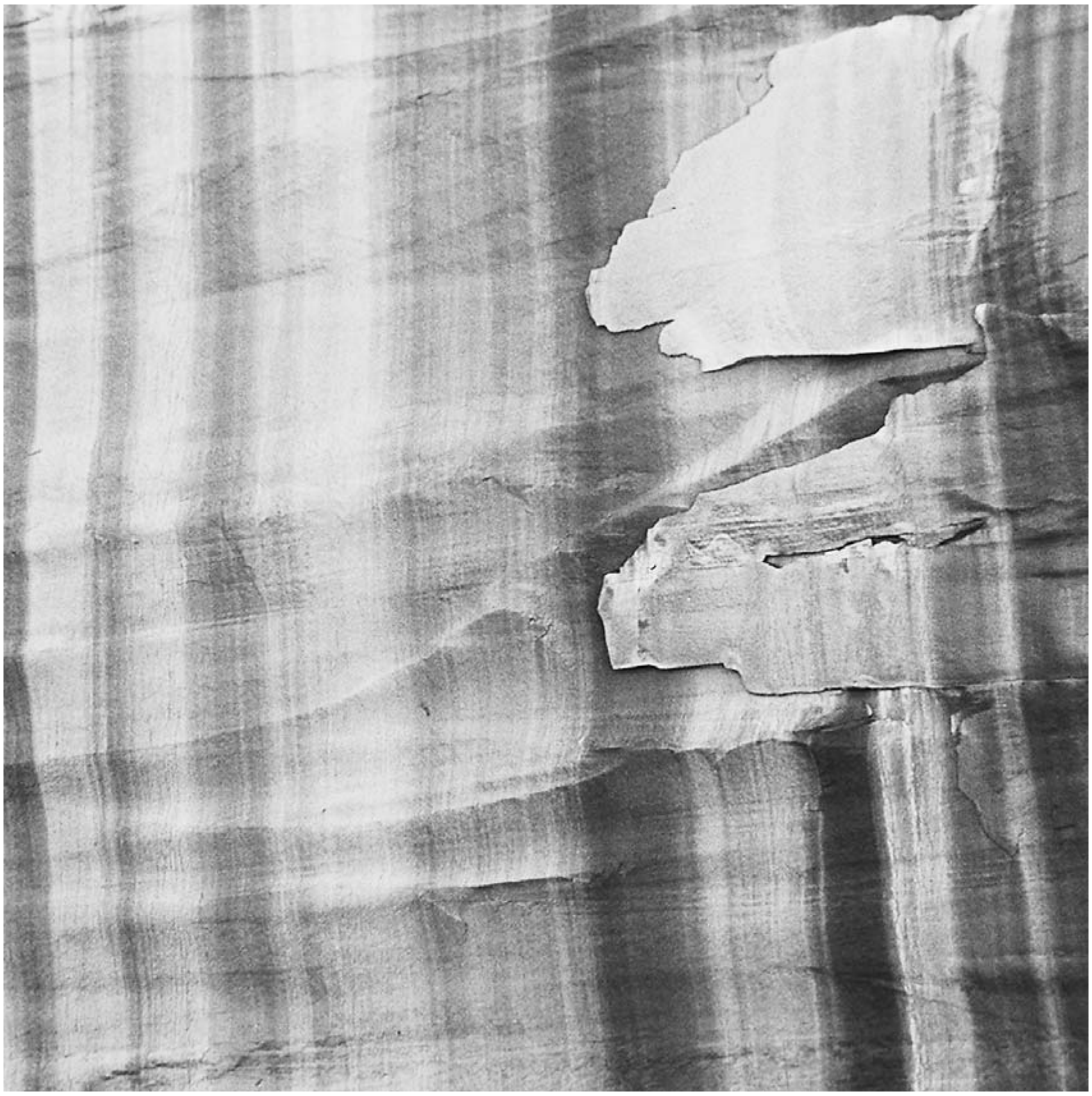
Historical geology, the study of Earth's physical history, is one of the two principal branches of geology, the other being physical geology, or the study of Earth's physical components and the forces that have shaped them. Among the principal subdisciplines of historical geology is stratigraphy, the study of rock layers, which are called strata or, in the singular form, a *stratum.*

Other important subdisciplines include geochronology, the study of Earth's age and the dating of specific formations in terms of geologic time; sedimentology, the study and interpretation of sediments, including sedimentary processes and formations; paleontology, the study of fossilized plants and animals; and paleoecology, the study of the relationship between prehistoric plants and animals and their environments. Several of these subjects are examined in other essays within this book.

EARLY WORK IN STRATIGRAPHY. Among the earliest contributions to what could be called historical geology came from the Italian scientist and artist Leonardo da Vinci (1452–1519), who speculated that fossils might have come from the remains of long-dead animals. Nearly two centuries later, stratigraphy itself had its beginnings when the Danish geologist Nicolaus Steno (1638–1687) studied the age of rock strata.

Steno formulated what came to be known as the law of superposition, or the idea that strata are deposited in a sequence such that the deeper the layer, the older the rock. This, of course, assumes that the rock has been undisturbed, and it is applicable only for one of the three major types of rock, sedimentary (as opposed to igneous or metamorphic). Later, the German geologist Johann Gottlob Lehmann (1719–1767) put forward the theory that certain groups of rocks tend to be associated with each other and that each layer of rock is a sort of chapter in the history of Earth.

Thus, along with Steno, Lehmann helped pioneer the idea of the stratigraphic column, discussed later in this essay. The man credited as the "father of stratigraphy," however, was the English engineer and geologist William Smith (1769–1839). In 1815 Smith produced the first modern geologic map, showing rock strata in England and Wales. Smith's achievement, discussed in Measuring and Mapping Earth, influenced all of geology to the present day by introducing the idea of geologic, as opposed to geographic, map-



STRIATIONS VISIBLE IN SANDSTONE FROM NEON CANYON, UTAH. *(© Rod Planck/Photo Researchers. Reproduced by permission.)*

ping. Furthermore, by linking stratigraphy with paleontology, he formulated an important division of stratigraphy, known as biostratigraphy.

AREAS OF STRATIGRAPHIC STUDY

Along with biostratigraphy, the major areas of stratigraphy include lithostratigraphy, chronostratigraphy, geochronometry, and magnetostratigraphy. The most basic type of stratigraphy, and the first to emerge, was lithostratigraphy, which is simply the study and description of rock layers. Earth scientists working in the area of lithostratigraphy identify various types of layers, which include (from the most specific to the most general), formations, members, beds, groups, and supergroups.

Biostratigraphy involves the study of fossilized plants and animals to establish dates for and correlate relations between stratigraphic layers. Scientists in this field also identify categories of biostratigraphic units, the most basic being a biozone. Magnetostratigraphy is based on the investigation of geomagnetism and the reversals in Earth's magnetic field that have occurred over time. (See Geomagnetism as well as the discussion of paleomagnetism in Plate Tectonics.)

Chronostratigraphy is devoted to studying the ages of rocks and what they reveal about geologic time, or the vast stretch of history (approximately 4.6 billion years, abbreviated 4.6 Ga) over which Earth's geologic development has occurred. It is concerned primarily with relative dating, whereas geochronometry includes the determination of absolute dates and time intervals. This typically calls for the use of radiometric dating.

The Stratigraphic Column

The stratigraphic column is the succession of rock strata laid down over the course of time, each of which correlates to specific phases in Earth's geologic history. The record provided by the stratigraphic column is most reliable for studying the Phanerozoic, the current eon of geologic history, as opposed to the Precambrian, which constituted the first three eons and hence the vast majority of Earth's geologic history. The relatively brief span of time since the Phanerozoic began (about 545 million years, or Ma) has seen by far the most dramatic changes in plant and animal life. It was in this eon that the fossil record emerged, giving us far more detailed information about comparatively recent events than about a much longer span of time in the more distant past.

RELATIVE AND ABSOLUTE DATING. Precambrian time is so designated because it precedes the Cambrian period, one of 11 periods in the Phanerozoic eon. The Cambrian period extended for about 50 million years, from approximately 545 Ma to 495 Ma ago. This statement in terms of years, however inexact, is an example of absolute age. By contrast, if we say that the Cambrian period occurred at the beginning of the Paleozoic era, after the end of the Proterozoic eon and before the beginning of the Ordovician period, this is a statement of relative age. Both statements are true, and though it is obviously preferable to measure time in absolute terms, sometimes relative terms are the only ones available.

Dating, in scientific terms, is any effort directed toward finding the age of a particular item or phenomenon. Relative dating methods assign an age relative to that of other items, whereas absolute dating determines age in actual years or millions of years. When geologists first embarked on stratigraphic studies, the only means of dating available to them were relative. Using Steno's law of superposition, they reasoned that a deeper layer of sedimentary rock was necessarily older than a shallower layer.

Advances in our understanding of atomic structure during the twentieth century, however, made possible a particularly useful absolute form of dating through the study of radioactive decay. Radiometric dating, which is explained in more detail in Geologic Time, uses ratios between "parent" and "daughter" isotopes. Radioactive isotopes decay, or emit particles, until they become stable, and as this takes place, parent isotopes spawn daughters. The amount of time that it takes for half the isotopes in a sample to stabilize is termed a *half-life.* Elements such as uranium, which has isotopes with half-lives that extend into the billions of years, make possible the determination of absolute dates for extremely old geologic materials.

DIVISIONS OF THE STRATIGRAPHIC COLUMN. Geologic time is divided into named groupings according to six basic units, which are (in order of size from longest to shortest) eon, era, period, epoch, age, and chron. There is no absolute standard for the length of any unit; rather, it takes at least two ages to make an epoch, at least two epochs to compose a period, and so on. The dates for specific eons, eras, periods, and so on are usually given in relative terms, however; an example is the designation of the Cambrian period given earlier.

Chronostratigraphy also uses six time units: the eonothem, erathem, system, series, stage, and chronozone. These time units are analogous to the terms in the geologic time scale, the major difference being that chronostratigraphic units are conceived in terms of relative time and are not assigned dates. The more distant in time a particular unit is, the more controversy exists regarding its boundary with preceding and successive units. This is true both of the geologic and the chronostratigraphic scales.

For this reason, the International Union of Geological Sciences, the leading worldwide body of geologic scientists, has established a Commission on Stratigraphy to determine such boundaries. The commission selects and defines what are called Global Stratotype Sections and Points (GSGPs), which are typically marine fossil formations. Because it is believed that life has existed longest on Earth in its oceans, samples from the water provide the most reliable stratigraphic record.

Naming of Chronostratigraphic Units

As noted, the chronostratigraphic divisions correspond to units of geologic time, even though chronostratigraphic units are based on relative dating methods and geologic ones use absolute time measures. Because attempts at relative dating have been taking place since the late eighteenth century, today's geologic units originated as what would be called *stratigraphic* or *chronostratigraphic* units. Even today the names of the phases are the same, with the only difference being the units in which they are expressed. Thus, when speaking in terms of geologic time, one would refer to the Jurassic period, whereas in stratigraphic terms, this would be the Jurassic *system.*

In 1759 the Italian geologist Giovanni Arduino (1714–1795) developed the idea of primary, secondary, and tertiary groups of rocks. Though the use of the terms *primary* and *secondary* has been discarded, vestiges of Arduino's nomenclature survive in the modern designation of the Tertiary subera of the Cenozoic era (erathem in stratigraphic terminology) as well as in the name of the present period or system, the Quaternary. (Just as primary, secondary, and tertiary refer to a first, second, and third level, respectively, the term *quaternary* indicates a fourth level.)

We are living in the fourth of four eons, or eonothems, the Phanerozoic, which is divided into three eras, or erathems: Paleozoic, Mesozoic, and Cenozoic. These eras, in turn, are divided into 11 periods, or systems, whose names (except for Tertiary and Quaternary) refer to the locations in which the respective stratigraphic systems were first observed. The names of these systems, along with their dates in millions of years before the present and the origin of their names, are as follows (from the most distant to the most recent):

<u>Periods/Systems of the Paleozoic Era/Erathem</u>

- Cambrian (about 545 to 495 Ma): Cambria, the Roman name for the province of Wales
- Ordovician (about 495 to 443 Ma): Ordovices, the name of a Celtic tribe in ancient Wales
- Silurian (about 443 to 417 Ma): Silures, another ancient Welsh Celtic tribe
- Devonian (about 417 to 354 Ma): Devonshire, a county in southwest England
- Mississippian (a subperiod of the Carboniferous period, about 354 to 323 Ma): the Mississippi River
- Pennsylvanian (a subperiod of the Carboniferous, about 323 to 290 Ma): the state of Pennsylvania
- Permian (about 290 to 248.2 Ma): Perm, a province in Russia

<u>Periods/Systems of the Mesozoic Era/Erathem</u>

- Triassic (about 248.2 to 205.7 Ma): a tripartite, or threefold, division of rocks in Germany
- Jurassic (about 205.7 to 142 Ma): the Jura Mountains of Switzerland and France
- Cretaceous (about 142 to 65 Ma): from a Latin word for "chalk," a reference to the chalky cliffs of southern England and France

Within the more recent Cenozoic era, or erathem, names of epochs (or "series" in stratigraphic terminology) become important. They are all derived from Greek words, whose meanings are given below:

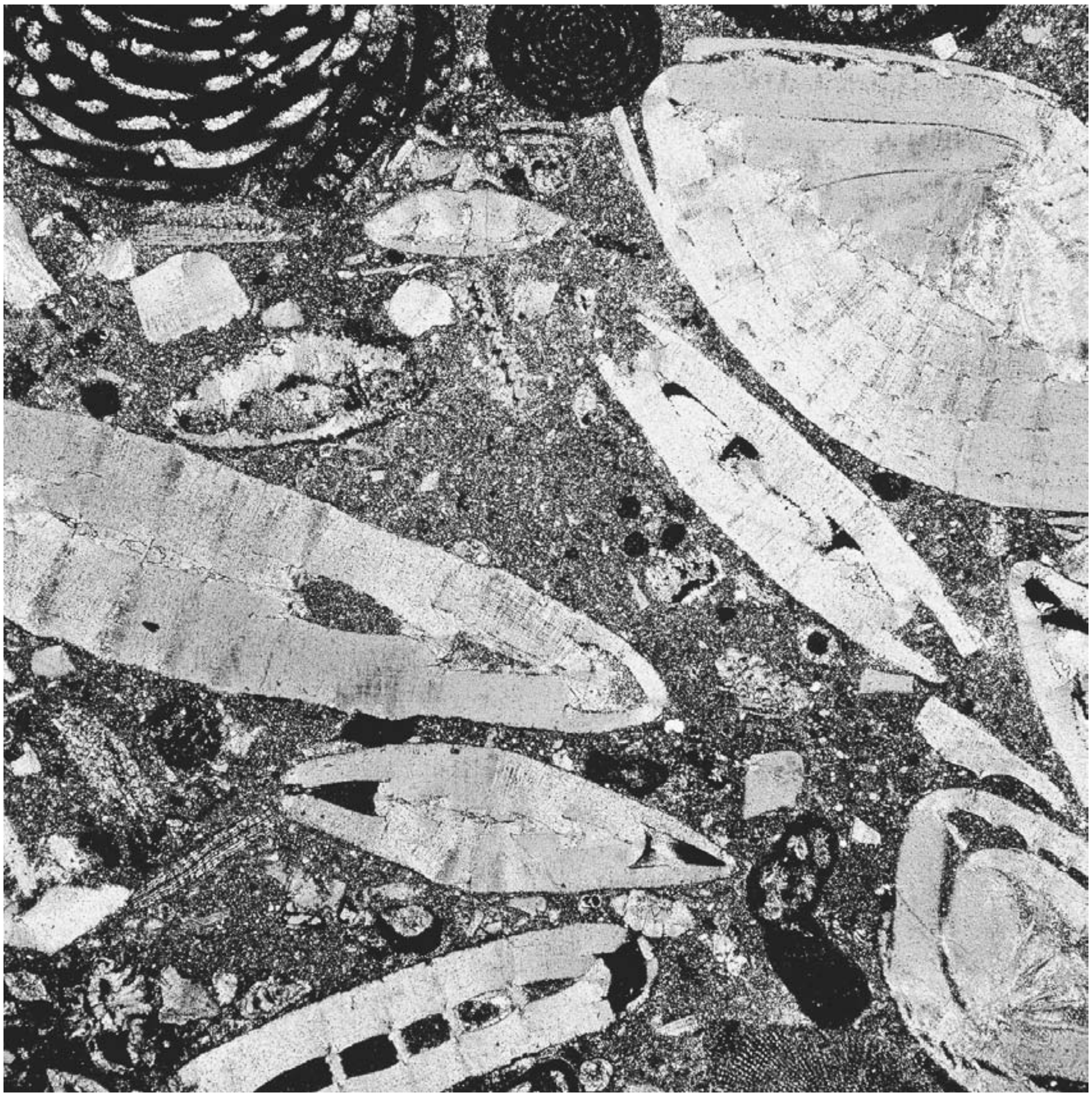
<u>Epochs/Series of the Cenozoic Era/Erathem</u>

- Paleocene (about 65 to 54.8 Ma): "early dawn of the recent"
- Eocene (about 54.8 to 33.7 Ma): "dawn of the recent"
- Oligocene (about 33.7 to 23.8 Ma): "slightly recent"
- Miocene (about 23.8 to 5.3 Ma): "less recent"
- Pliocene (about 5.3 to 1.8 Ma): "more recent"
- Pleistocene (about 1.8 to 0.01 Ma): "most recent"
- Holocene (about 0.01 Ma to present): "wholly recent"

REAL-LIFE APPLICATIONS

CORRELATION

The geologist studying the stratigraphic record is a sort of detective, looking for clues. Just as detectives have their methods for solving crimes, geologists rely on correlation, or methods of establishing age relationships between various strata. There are two basic types of correlation: physical correlation, which requires comparison of the physical characteristics of the strata, and fossil correlation, the comparison of fossil types.



A POLARIZED-LIGHT MICROGRAPH SHOWS FOSSILS IN LIMESTONE DATING TO THE EOCENE AND OLIGOCENE EPOCHS. *(© A. Pasieka/Photo Researchers. Reproduced by permission.)*

Actually, chronostratigraphic work is very similar some of the toughest cases confronted by police detectives, because more often than not the geologic detective has little evidence on which to operate. First of all, as noted earlier, only sedimentary rock can be used in making such determinations: for instance, igneous rock in its molten form, as when it is expelled from a volcano, could force itself underneath a rock stratum, thus confusing the stratigraphic record.

POTENTIAL PITFALLS. Even when the rock is sedimentary, there is still plenty of room for error. The layers may be many feet or less than an inch deep, and it is up to the geologist to determine whether the stratum has been affected by such geologic forces as erosion. If erosion has occurred, it can cause a disturbance, or unconformity (discussed later), which tends to render inaccurate any reading of the stratigraphic record.

Another possible source of disturbance is an earthquake, which could cause one part of Earth's crust to shift over an adjacent section, making the stratigraphic record difficult, if not impossible, to read. Under the best of conditions, after all, the strata are hardly neat, easily defined lines. If one observes a horizontal section, there is likely to be a change in thickness, because as the stratum extends outward, it merges with the edges of adjacent deposits.

Yet another potential pitfall in stratigraphic correlation involves one of the most useful tools

KEY TERMS

ABSOLUTE AGE: The absolute age of a geologic phenomenon is its age in Earth years. Compare with *relative age.*

BIOSTRATIGRAPHY: An area of stratigraphy involving the study of fossilized plants and animals in order to establish dates for and correlations between stratigraphic layers.

CHRONOSTRATIGRAPHY: A subdiscipline of stratigraphy devoted to studying the relative ages of rocks. Compare with *geochronometry.*

CORRELATION: A method of establishing age relationships between various rock strata. There are two basic types of correlation: physical correlation, which requires comparison of the physical characteristics of the strata, and fossil correlation, the comparison of fossil types.

DATING: Any effort directed toward finding the age of a particular item or phenomenon. Methods of geologic dating are either relative (i.e., comparative and usually based on rock strata) or absolute. The latter, based on such methods as the study of radioactive isotopes, usually is given in terms of actual years or millions of years.

EON: The longest phase of geologic time, equivalent to an eonothem in the stratigraphic time scale. Earth's history has consisted of four eons, the Hadean or Priscoan, Archaean, Proterozoic, and Phanerozoic. The next-smallest subdivision of geologic time is the era.

EPOCH: The fourth-longest phase of geologic time, shorter than an era and longer than an age and a chron. An epoch is equivalent to a series in the stratigraphic time scale. The current epoch is the Holocene, which began about 0.01 Ma (10,000 years) ago.

ERA: The second-longest phase of geologic time, after an eon, and equivalent to an erathem in the stratigraphic time scale. The current eon, the Phanerozoic, has had three eras, the Paleozoic, Mesozoic, and Cenozoic, which is the current era. The next-smallest subdivision of geologic time is the period.

EROSION: The movement of soil and rock due to forces produced by water, wind, glaciers, gravity, and other influences.

GA: An abbreviation meaning "giga-years" or "billion years." The age of Earth is about 4.6 Ga.

GEOCHRONOMETRY: An area of stratigraphy devoted to determining absolute dates and time intervals. Compare with *chronostratigraphy.*

GEOLOGIC MAP: A map showing the rocks beneath Earth's surface, including their distribution according to type as well as their ages, relationships, and structural features.

GEOLOGIC TIME: The vast stretch of time over which Earth's geologic development has occurred. This span (about 4.6 billion years) dwarfs the history of human existence, which is only about two million years. Much smaller still is the span of human civilization, only about 5,500 years.

KEY TERMS CONTINUED

HISTORICAL GEOLOGY: The study of Earth's physical history. Historical geology is one of two principal branches of geology, the other being physical geology.

ISOTOPES: Atoms that have an equal number of protons, and hence are of the same element, but differ in their number of neutrons. This results in a difference of mass. An isotope may be either stable or radioactive.

LAW OF FAUNAL SUCCESSION: The principle that all samples of any given fossil species were deposited on Earth, regardless of location, at more or less the same time. This makes it possible to correlate widely separated strata.

LAW OF SUPERPOSITION: The principle that strata are deposited in a sequence such that the deeper the layer, the older the rock. This is applicable only or sedimentary rock, as opposed to igneous or metamorphic rock.

LITHOSTRATIGRAPHY: An area of stratigraphy devoted to the study and description (but not the dating) of rock layers.

MA: An abbreviation used by earth scientists, meaning "million years" or "megayears." When an event is designated as, for instance, 160 Ma, it usually means 160 million years ago.

PALEONTOLOGY: The study of fossilized plants and animals, or flora and fauna.

PERIOD: The third-longest phase of geologic time, after an era; it is equivalent to a system in the stratigraphic time scale. The current eon, the Phanerozoic, has had 11 periods, and the current era, the Cenozoic, has consisted of three periods, of which the most recent is the Quaternary. The next-smallest subdivision of geologic time is the epoch.

PRECAMBRIAN TIME: A term that refers to the first three of four eons in Earth's history, which lasted from about 4,560 Ma to about 545 Ma ago.

RADIOACTIVITY: A term describing a phenomenon whereby certain materials are subject to a form of decay brought about by the emission of high-energy particles or radiation. Forms of particles or energy include alpha particles (positively charged helium nuclei), beta particles (either electrons or subatomic particles called positrons), or gamma rays, which occupy the highest energy level in the electromagnetic spectrum.

RADIOMETRIC DATING: A method of absolute dating using ratios between "parent" isotopes and "daughter" isotopes, which are formed by the radioactive decay of parent isotopes.

RELATIVE AGE: The relative age of a geologic phenomenon is its age compared with the ages of other geologic phenomena, particularly the stratigraphic record of rock layers. Compare with *absolute age.*

KEY TERMS CONTINUED

SEDIMENT: Material deposited at or near Earth's surface from a number of sources, most notably preexisting rock.

SEDIMENTARY ROCK: Rock formed by compression and deposition (i.e., formation of deposits) on the part of other rock and mineral particles. Sedimentary rock is one of the three major types of rock, along with igneous and metamorphic.

SEDIMENTOLOGY: The study and interpretation of sediments, including sedimentary processes and formations.

STRATA: Layers, or beds, of rocks beneath Earth's surface. The singular form is *stratum.*

STRATIGRAPHIC COLUMN: The succession of rock strata laid down over the course of time, each of which correlates to specific junctures in Earth's geologic history.

STRATIGRAPHY: The study of rock layers, or strata, beneath Earth's surface.

UNCONFORMITY: An apparent gap in the geologic record, as revealed by observing rock layers or strata.

WEATHERING: The breakdown of rocks and minerals at or near the surface of Earth due to physical or chemical processes, or both.

available to a geologist attempting to find an absolute age for the materials he or she is studying: radiometric dating. Though this method can provide accurate absolute dates, it is quite possible that the age thus determined will be the age of the parent rock from which a sample is taken, not the age of the sample itself. The grains of sand in a piece of sandstone, for instance, are much older than the larger unit of sandstone, and for this reason, radiometric dating is useful only in specific circumstances.

PHYSICAL AND FOSSIL CORRELATION. Given all these challenges, it is a wonder that geologists manage to correlate strata successfully, yet they do. Physical correlations are achieved on the basis of several criteria, including color, the size of grains, and the varieties of minerals found within a stratum. By such means, it is sometimes possible to correlate widely separated strata.

Particularly impressive feats of correlation can result from the study of fossils, whose stratigraphic implications, as we have noted, were first discovered by William Smith. Smith hit upon the idea of biostratigraphy while excavating land for a set of canals near London. As he discovered, any given stratum contains the same types of fossils, and strata in two different areas thus can be correlated.

Long before his countryman Charles Darwin (1809–1882) developed the theory of evolution, Smith conceived his own law of faunal succession, which hints at the idea that species developed and disappeared over given phases in Earth's past. According to the law of faunal succession, all samples of any given fossil species were deposited on Earth, regardless of location, at more or less the same time. As a result, if a geologist finds a stratum in one area that contains a particular fossil and another in a distant area containing the same fossil, it is possible to conclude that the strata are the same.

Unconformities

In discussing the many challenges facing a geologist studying stratigraphic data, the role of erosion was noted. Let us return to that subject, because erosion is a source of what are known as unconformities, or gaps in the rock record. Unconformities are of three types: angular unconformities, disconformities, and nonconformities.

Angular unconformities involve a tilting of the layers, such that an upper layer does not lie perfectly parallel to a lower one. Disconformities are more deceptive, because the layers are parallel, yet there is still an unconformity between them, and only a study of the fossil record can reveal the unconformity. Finally, a nonconformity arises when sedimentary rocks are divided from a type of igneous rock known as *intrusive* (meaning "cooled within Earth").

ANGULAR UNCONFORMITIES. Angular unconformities emerge as a by-product of the dramatic shifts and collisions that take place in plate tectonics (see Plate Tectonics). Sediment accumulates and then, as a result of plate movement, is moved about and eventually experiences weathering and erosion. Layers are tilted and then flattened by more erosion, and as the solid earth rises or sinks, they are shifted further. Such is the case, for instance, along the Colorado River at the Grand Canyon, where angular unconformities reveal a series of movements over the years.

Another famous angular unconformity can be found at Siccar Point in Scotland, where nearly horizontal deposits of sandstone rest atop nearly vertical ones of graywacke, another sedimentary rock. Observations of this unconformity led the great geologist James Hutton (1726–1797) to the realization that Earth is much, much older than the 6,000 years claimed by theologians in his day (see Historical Geology).

WHERE TO LEARN MORE

Bishop, A. C., A. Woolley, and A. Hamilton. *Cambridge Guide to Minerals, Rocks, and Fossils*. New York: Cambridge University Press, 1992.

Boggy's Links to Stratigraphy and Geochronology (Web site). <http://geologylinks.freeyellow.com/stratigraphy.html>.

Harris, Nicholas, Alessandro Rabatti, and Andrea Ricciardi. *The Incredible Journey to the Beginning of Time.* New York: Peter Bedrick Books, 1998.

Lamb, Simon, and David Sington. *Earth Story: The Shaping of Our World.* Princeton, NJ: Princeton University Press, 1998.

MacRae, Andrew. *Radiometric Dating and the Geological Time Scale* (Web site). <http://www.talkorigins.org/faqs/dating.html>.

Reeves, Hubert. *Origins: Cosmos, Earth, and Mankind.* New York: Arcade, 1998.

Spickert, Diane Nelson, and Marianne D. Wallace. *Earthsteps: A Rock's Journey Through Time.* Golden, CO: Fulcrum Kids, 2000.

Stratigraphy and Earth History—West's Geology Directory (Web site). <http://www.soton.ac.uk/~imw/stratig.htm>.

University of Georgia Stratigraphy Lab (Web site). <http://www.uga.edu/~strata/home.html>.

Web Time Machine. UCMP (University of California, Berkeley, Museum of Paleontology) (Web site). <http://www.ucmp.berkeley.edu/help/timeform.html>.

PALEONTOLOGY

CONCEPT

Thanks to a certain 1993 blockbuster, most people know the name of at least one period in geologic history. *Jurassic Park* spurred widespread interest in dinosaurs and, despite its fantastic plot, encouraged popular admiration and respect for the work of paleontologists. Paleontology is the study of life-forms from the distant past, as revealed primarily through the record of fossils left on and in the earth. It is a complex and varied subdiscipline of historical geology that is tied closely to the biological sciences. As with other types of historical geology, the work of a paleontologist is similar to that of a detective investigating a case with few and deceptive clues. Reliable fossil samples are more rare, compared with the vast number of species that have lived on Earth, than one might imagine. Furthermore, several factors pose challenges for paleontologists attempting to interpret the fossil record. Nonetheless, paleontologic research has led to a growing understanding of how life emerged, how Earth has changed, and how vast animal populations became extinct over relatively short periods of time.

HOW IT WORKS

Thinking in Terms of Geologic Time

The term geologic time refers to the great sweep of Earth's history, a timescale that dwarfs the span of human existence. The essays Historical Geology and Geologic Time offer several comparisons to emphasize the proportions involved and to illustrate the very short period during which human life has existed on this planet.

As one example shows, if all of geologic time were compressed into a single year, the first *Homo sapiens* would have appeared on the scene at about 8:00 p.m. on December 31. Human civilization, which dates back about 5,500 years (a millennium before the building of Egypt's great pyramids) would have emerged within the last minute of the year.

In another example, geologic time is compared to the distance from Los Angeles to New York City. On this scale, the period of time in which humans have existed on the planet would be equivalent to the distance from New York's Central Park to the Empire State Building, or less than 2 mi. (3.2 km). The history of human civilization, on the other hand, would be less than 16 ft. (4.9 m) long.

THE "ABYSS OF TIME." Needless to say, the scope of geologic time compared with the units with which we are accustomed to measuring our lives (or even the history of our civilization) is more than a little intimidating. This fact perhaps was best expressed by the Scottish geologist John Playfair (1748–1819), friend and countryman of the "father of geology," James Hutton (1726–1797). At a time when many people were content to believe that the Earth had been around no more than 6,000 years (see Historical Geology), Hutton suggested that to undergo the complex processes that had shaped its landforms, the planet had to be much, much older. Commenting on Hutton's discoveries, Playfair said, "The mind seemed to grow giddy by looking so far into the abyss of time."

THE HEAD OF A TYRANNOSAURUS REX ("KING OF THE TERRIBLE LIZARDS") FOUND IN MONTANA, DATING TO THE MESOZOIC PERIOD. *(© Tom McHugh/Photo Researchers. Reproduced by permission.)*

CARBON: THE MEANING OF "LIFE"

A discussion of life on Earth requires us to go deep into this "abyss," though not nearly as far back as the planet's origins. It does appear that life on Earth existed at a very early point, but in this context "life" refers merely to molecules of carbon-based matter capable of replicating themselves. Knowledge of these very early forms is extremely limited.

Carbon appears in all living things, in things that were once living, and in materials produced by living things (for example, sap, blood, and urine). Hence, the term *organic,* which once meant only living matter, refers to almost all types of material containing carbon. The only carbon-containing materials that are not considered organic are oxides, such as carbon dioxide and carbon monoxide, and carbonates, a class of minerals that is extremely abundant on Earth.

PRECAMBRIAN TIME

We will return to the subject of carbon, which plays a role in one technique for dating relatively recent items or phenomena. For the present, however, let us set our bearings for a discussion of the Phanerozoic eon, the fourth and last of the major divisions of geologic time. Though extremely primitive life-forms existed before the Phanerozoic eon, the vast majority of species have evolved since it began, and consequently paleontological work is concerned primarily with the Phanerozoic eon.

The divisions of geologic time are not arranged in terms of strict mathematical relationships of the type to which we are accustomed, for example, ten years in a decade, ten decades in a century, and so on. Instead, each era consists of two or more periods, each period consists of two or more epochs, and so on. The first 4,000 million years or so of Earth's existence (abbreviated as 4,000 Ma, or 4 Ga) are known as Precambrian time. In discussing this period of time, the vast majority of the planet's history, it is seldom necessary to speak of geologic time divisions smaller than the largest unit, the eon. Precambrian time consisted of three eons, the Hadean or Priscoan, Archaean, and Proterozoic.

THE FIRST THREE EONS. The Hadean (sometimes called the Priscoan and dating to about 4,560 Ma to 4,000 Ma ago) saw the formation of the planet and the beginnings of the oceans and an early form of atmosphere that consisted primarily of carbon dioxide. It was during this eon that the carbon-based matter referred earlier made its appearance, perhaps by means of the meteorites that bombarded the planet during that long-ago time.

In the Archaean eon (about 4,000 Ma to 2,500 Ma ago) the first clear evidence of life appeared in the form of microorganisms. These were prokaryotes, or cells without a nucleus, which eventually were followed by eukaryotes, or cells with a nucleus. Many of the prerequisites for life as we know it were established during this time, though our present oxygen-containing atmosphere still lay far in the future.

Longest of the four eons was the Proterozoic eon (about 2,500 Ma to 545 Ma). This phase saw the beginnings of very basic forms of plant life, while oxygen in the atmosphere assumed about 4% of its present levels. Animal life, meanwhile, still consisted primarily of eukaryotes.

The Phanerozoic Eon

The majority of paleontologic history has taken place during the Phanerozoic eon. In the course of this essay, we discuss its eras and periods (the second- and third-longest spans of geologic time, respectively) as they relate to life on Earth. The three Phanerozoic eras are as follows:

Eras of the Phanerozoic Eon

- Paleozoic (about 545–248.2 Ma)
- Mesozoic (about 248.2–65 Ma)
- Cenozoic (about 65 Ma–present)

Within these eras are the following periods:

Periods of the Paleozoic Era

- Cambrian (about 545–495 Ma)
- Ordovician (about 495–443 Ma)
- Silurian (about 443–417 Ma)
- Devonian (about 417–354 Ma)
- Carboniferous (about 354–290 Ma)
- Permian (about 290–248.2 Ma)

Periods of the Mesozoic Era

- Triassic (about 248.2–205.7 Ma)
- Jurassic (about 205.7–142 Ma)
- Cretaceous (about 142–65 Ma)

Periods of the Cenozoic Era

- Palaeogene (about 65–23.8 Ma)
- Neogene (about 23.8–1.8 Ma)
- Quaternary (about 1.8 Ma to the present)

The Carboniferous period of the Paleozoic era usually is divided into two subperiods, the Mississippian (about 354 to 323 Ma) and the Pennsylvanian (about 323–290 Ma). In addition, the Palaeogene and Neogene periods of the Cenozoic era often are lumped together as a subera called the Tertiary. By substituting that name for those of the two periods, it is possible to use a time-honored mnemonic device by which geology students have memorized the names of the 11 Phanerozoic periods: "Camels Ordinarily Sit Down Carefully; Perhaps Their Joints Creak Tremendously Quietly."

EPOCHS OF THE CENOZOIC ERA. An epoch is the fourth-largest division of geologic time and is, for the most part, the smallest one with which we will be concerned. (There are two smaller categories, the age and the chron.) Listed here are the epochs of the Cenozoic era from the most distant to the Holocene, in which we are now living. Their names are derived from Greek words whose meanings are provided:

Epochs of the Cenozoic Era

- Paleocene (about 65–54.8 Ma): "early dawn of the recent"
- Eocene (about 54.8–33.7 Ma): "dawn of the recent"
- Oligocene (about 33.7–23.8 Ma): "slightly recent"
- Miocene (about 23.8–5.3 Ma): "less recent"
- Pliocene (about 5.3–1.8 Ma): "more recent"
- Pleistocene (about 1.8–0.01 Ma): "most recent"
- Holocene (about 0.01 Ma to present): "wholly recent"

A Brief Overview of Paleontologic History

The title "A Brief Overview of Paleontologic History" is almost a contradiction in terms, since virtually nothing about the history of Earth has been brief. Moreover, the history of *life* on Earth is so filled with detail and complexity that it could fill many books, as indeed it has. Owing to that complexity, anything approaching an exhaustive treatment of the subject would burden the reader with so much technical terminology that it would obscure the larger overview of paleontology and the materials of the paleontologist's work. Therefore, only the most cursory of treatments is possible, or indeed warranted, in the present context. For additional detail, the reader is invited to consult other texts, including

A DINOSAUR IS EXCAVATED AT DINOSAUR NATIONAL MONUMENT IN COLORADO. *(© James L. Amos/Photo Researchers. Reproduced by permission.)*

those listed in the suggested reading section at the end of this essay.

As with many another process, the evolution of organisms was exceedingly slow in the beginning (and here the comparative term *slow* refers even to the standards of geologic time), but it sped up considerably over the course of Earth's history. This is not to suggest that the development of life-forms has been a steady process; on the contrary, it has been punctuated by mass extinctions, discussed at the conclusion of this essay. Nonetheless, it is correct to say that during the first 80%–90% of Earth's history, the few existing life-forms underwent an extremely slow process of change.

PRECAMBRIAN AND PALEOZOIC LIFE-FORMS. Life existed in Precambrian time, as noted, but over the course of those four billion years, it evolved only to the level of single-cell microorganisms. Samples of these organisms have been found in the fossil record, but the fossilized history of life on Earth really began in earnest only with the Cambrian period at the beginning of the Paleozoic era and the Phanerozoic eon. The early Cambrian period saw an explosion of invertebrate (without an internal skeleton) marine forms, which dominated from about 545 Ma–417 Ma ago. By about 420 Ma–410 Ma, life had appeared on land, in the form of algae and primitive insects.

The beginning of the Devonian period (approximately 417 Ma) saw the appearance of the first vertebrates (animals with an internal skeleton), which were jawless fish. Plant life on land consisted of ferns and mosses. By the late Devonian (about 360 Ma), fish had evolved jaws, and amphibians had appeared on land. Reptiles emerged between about 320 Ma and 300 Ma, in the Pennsylvanian subperiod of the Carboniferous. In the last period of the Paleozoic era, the Permian (about 290–248.2 Ma), reptiles became the dominant land creatures.

MESOZOIC AND CENOZOIC LIFE-FORMS. The next era, the Mesozoic (about 248.2–65 Ma), belonged to a particularly impressive form of reptile, known as the terrible lizard: the dinosaur. These creatures are divided into groups based on the shape of their hips, which were either lizardlike or birdlike. Though the lizardlike Saurischia emerged first, they lived alongside the birdlike Ornithischia throughout the late Triassic, Jurassic, and Cretaceous periods. Ornithischia were all herbivores, or plant eaters, whereas Saurischia included both herbivores and carnivores, or meat eaters. Naturally, the most fierce of the dinosaurs were carnivores, a group that included the largest carnivore ever to walk the earth, *Tyrannosaurus.*

Though dinosaurs receive the most attention, the Mesozoic world was alive with varied forms, including flying reptiles and birds. (In fact, dinosaurs may have been related to birds, and, in the opinion of some paleontologists, they may have been warm-blooded, like birds and mammals, rather than cold-blooded, like other reptiles.) Botanical life included grasses, flowering plants, and trees of both the deciduous (leaf-shedding) and coniferous (cone-bearing) varieties.

A violent event, discussed in the context of mass extinction later in this essay, brought an end to the Mesozoic era. This cleared the way for the emergence of mammalian forms at the beginning of the Cenozoic, though it still would be a long time before anything approaching an ape, let alone a human, appeared on the scene. The earliest hominid, or humanlike creature, dates back to about four million years ago, in the Pliocene epoch of the Neogene period.

Historical Geology and Paleontology

One of the two principal divisions of geology (along with physical geology) is historical geology, the study of Earth's physical history. Other subdisciplines of historical geology are stratigraphy, the study of rock layers, or strata, beneath Earth's surface; geochronology, the study of Earth's age and the dating of specific formations in terms of geologic time; and sedimentology, the study and interpretation of sediments, including sedimentary processes and formations.

Paleontology, the investigation of life-forms from the distant past (primarily through the study of fossilized plants and animals), is another subdiscipline of historical geology. Though it is rooted in the physical sciences, it obviously crosses boundaries into the biological or life sciences as well. Related or subordinate fields include paleozoology, which focuses on the study of prehistoric animal life; paleobotany, the study of past plant life; and paleoecology, the study of the relationship between prehistoric plants and animals and their environments.

Classifying Plants and Animals

Given the close relationship between paleontology and the biological sciences, it is necessary to discuss briefly the taxonomic system applied in biology, botany, zoology, and related fields. Taxonomy is an area of biology devoted to the identification, classification, and naming of organisms. Devised in the eighteenth century by the Swedish botanist Carolus Linnaeus (1707–1778) and improved in succeeding years by many others, the taxonomic system revolutionized biology.

Linnaeus's taxonomy provided a framework for classifying known species not simply by superficial similarities but also by systemic characteristics. For example, worms and snakes have something in common on a surface level, because they are both without appendages and move by writhing on the ground. A worm is an invertebrate, however, whereas a snake is a vertebrate. The Linnaean system therefore would classify them in widely separated categories: they are not siblings or even first cousins but more like fourth cousins.

Moreover, the system created by Linnaeus gave scientists a means for classifying and thereby potentially understanding much about the history and characteristics of species as yet undiscovered. Thus, it would prove of immeasurable significance to the English naturalist Charles Darwin (1809–1882) in formulating his theory of evolution. As Darwin showed, the varieties of different organisms have increased over time, as those organisms developed characteristics that made them more adaptable to their environments. Plants and animals that failed to adapt simply became extinct, though failure to adapt is only one of several causes for extinction, as we shall see.

A BRIEF OVERVIEW. The Linnaean system uses binomial nomenclature, or a two-part naming scheme (in Latin), to identify each separate type of organism. If a man is named John Smith, then "Smith" identifies his family, while John identifies him singularly. Likewise each variety of organism is identified by genus, equivalent to Smith, and species, analogous to John. In the Linnaean system, there are eight levels of classification, which, from most general to most specific, are kingdom, phylum, subphylum, class, order, family, genus, and species.

These levels can be illustrated by identifying a species near and dear to all of us: *Homo sapiens,* commonly known as *humans.* We belong to the animal kingdom (Animalia), the Chordata (i.e., possessing some form of central nervous system) phylum, and the Vertebrata subphylum, indicating the existence of a backbone. Within the mammal (Mammalia) class we are part of the primate (Primata) order, along with apes. Humans are distinguished further as members of the hominid (Hominoidea), or "human-like" family; the genus *Homo* ("man"); and the species *sapiens* ("wise").

Dating Materials From the Past

In studying the past, paleontologists and other earth scientists working in the field of historical geology rely on a variety of dating techniques. "Dating," in a scientific context, usually refers to any effort directed toward finding the age of a particular item or phenomenon. It may be relative, devoted to finding an item's age in relation to that of other items; or absolute, involving the determination of age in actual years or millions of years.

Among the methods of relative dating are stratigraphic dating, discussed in the essay Stratigraphy, as well as seriation, faunal dating, and pollen dating. Seriation entails analyzing the abundance of a particular item and assigning relative dates based on that abundance. Faunal dating is the use of bones from animals (fauna) to determine age, and pollen dating, or palynology, analyzes pollen deposits.

FAUNAL DATING AND PALYNOLOGY. The concept of faunal dating emerged from early work by the English engineer and geologist William Smith (1769–1839), widely credited as the "father of stratigraphy." In particular, Smith established an important division of stratigraphy, known as biostratigraphy, that is closely tied to paleontology. While excavating land for a set of canals near London, he discovered that any given stratum, or rock layer, contains the same types of fossils, and therefore strata in two different areas can be correlated.

Smith stated this in what became known as the law of faunal succession: all samples of any given fossil species were deposited on Earth, regardless of location, at more or less the same time. As a result, if a geologist finds a stratum in one area that contains a particular fossil and another in a distant area containing the same fossil, it is possible to conclude that the strata are the same.

Pollen dating, or palynology, is based on the fact that seed-bearing plants release large numbers of pollen grains each year. As a result, pollen spreads over the surrounding area, and in many cases pollen from the distant past has been preserved. This has occurred primarily in lake beds, peat bogs, and, occasionally, in areas with cool or acidic soil. By observing the species of pollen deposited in an area, scientists are able to develop a sort of "pollen calendar," which provides information about such details as changes in climate.

DENDROCHRONOLOGY. Scientists use relative dating when they must, but they would prefer to determine dates in an absolute sense wherever possible. Most methods of absolute dating rely on processes that are not immediately comprehensible to the average person, but there is one exception: dendrochronology, or the dating of tree rings. As almost everyone knows, trees produce one growth ring per year. There is nothing magical about this, since a year is not an abstract unit of time; rather, it is based on Earth's revolution around the Sun, during which time the planet undergoes changes in orientation that result in the four seasons, which, in turn, affect the tree's growth.

Though dendrochronology makes use of a principle familiar to most people, the work of the dendrochronologist requires detailed, often complex study. Just as the layers of rock beneath Earth's surface reveal information about past geologic events (a matter discussed in the essay Stratigraphy), tree rings can tell us much about environmental changes. Thin rings, for instance, suggest climatic anomalies and may provide clues about cataclysmic events that were understood only vaguely by the ancient humans who experienced them. (An example of this is the apparent cataclysm of A.D. 535, which is discussed Earth Systems.)

AMINO-ACID RACIMIZATION. Dendrochronology is useful only for studying the relatively recent past, up to about 10,000 years—a span equivalent to the Holocene epoch, which began with the end of the last ice age. To investigate more distant phases of Earth's history, it is necessary to use forms of radiometric dating, which we will discuss shortly. The principles of radiometric dating, however, are illustrated by another method, amino-acid racimization.

With the exception of some microbes, living organisms incorporate only one of two forms of amino acids, known as L-forms. Once the organism dies, the L-amino acids gradually convert to D-amino acids. In the 1960s, scientists discovered that by comparing the ratios between the L- and D-forms, it was possible to date organisms that were several thousand years old. Unfortunately, it has since come to light that because of the many factors affecting the rate of amino-acid conversion, this method is less reliable than once was believed. Moisture, temperature, and pH (the relative acidity and alkalinity of a substance) all play a part, and because these factors vary so widely, amino-acid racimization no longer is used commonly.

Nonetheless, the basic principle behind amino-acid racimization plays a part in other, more reliable forms of absolute dating. Many of them are based on the fact that over time, a particular substance converts to another, mirror substance. By comparing the ratios between them, it is possible to arrive at some estimate of the amount of time that has elapsed since the organism died.

RADIOCARBON DATING. The most significant method of absolute dating available to scientists today is radiometric dating, which is explained in detail in the essay Geologic Time. Each chemical element is distinguished by the number of protons (positively charged particles) in its atomic nucleus, but atoms of a particular element may have differing numbers of neutrons, or neutrally charged particles, in their nuclei. Such atoms are referred to as isotopes.

Certain isotopes are stable, whereas others are radioactive, meaning that they are likely to eject particles from the nucleus over time. The amount of time it takes for half the isotopes in a sample to stabilize is called its half-life. By analyzing the quantity of radioactive isotopes in a given sample that have converted to stable isotopes, it is possible to determine the age of the sample. In other situations, it is necessary to compare ratios of unstable "parent" isotopes to even more unstable "daughter" isotopes produced by the parent.

As we noted earlier, carbon is present in all living things, and thus an important means of dating available to paleontologists uses a radioactive form of carbon. All atoms of carbon have six protons, and the most stable and abundant carbon isotope is carbon-12, so designated because it has six neutrons. On the other hand, carbon-14, with eight neutrons, is unstable.

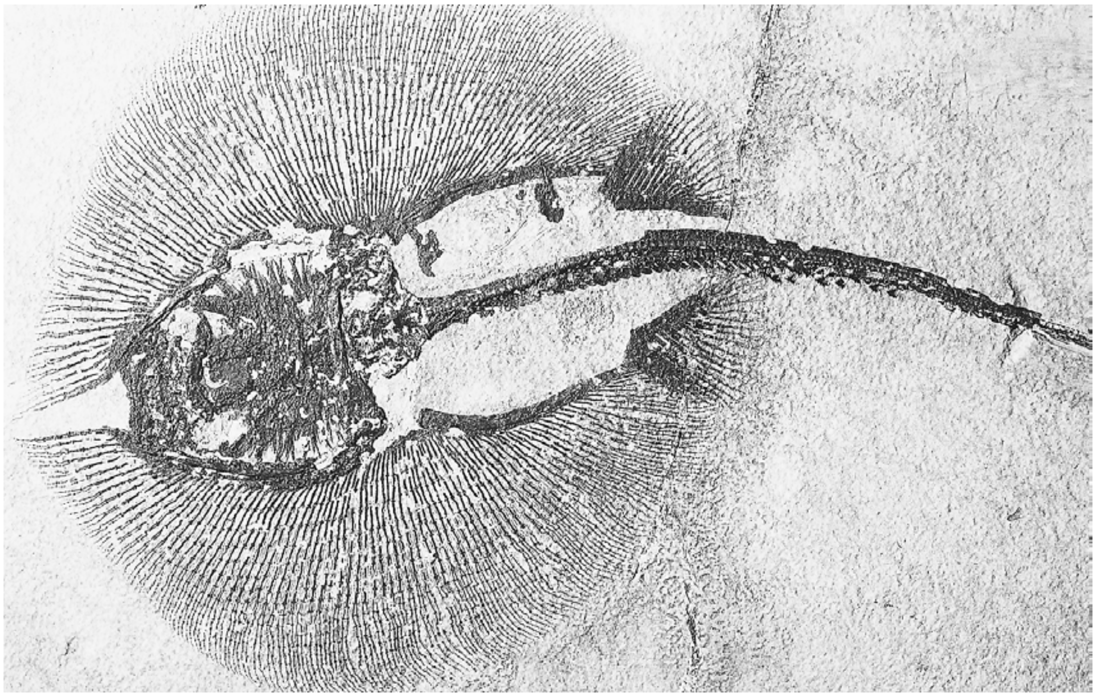
When an organism is alive, it incorporates a certain ratio of carbon-12 in proportion to the (very small) amount of carbon-14 that it receives from the atmosphere. Once the organism dies, however, it stops incorporating new carbon, and the ratio between carbon-12 and carbon-14 begins to change as the carbon-14 decays to form nitrogen-14. Therefore, a scientist can use the ratios of carbon-12, carbon-14, and nitrogen-14 to estimate the age of an organic sample. This method is known as radiocarbon dating.

Carbon-14, or radiocarbon, has a half-life of 5,730 years, meaning that it is useful for analyzing only fairly recent samples. Nonetheless, it takes much longer than 5,730 years for the other half of the radiocarbon isotopes in a given sample to stabilize, and for this reason radiocarbon dating can be used with considerable accuracy for 30,000–40,000 years. Sophisticated instrumentation can extend this range even further, up to 70,000 years.

THE LIMITS OF ABSOLUTE DATING. While 70,000 years, or 0.07 Ma, may be a long time in human terms, from the standpoint of the earth scientist, 0.07 Ma is only yesterday—the latter part of the last epoch, the Pleistocene. Other forms of radiometric dating, such as potassium-argon dating and uranium-series dating, can be used to measure truly long spans of times, in the billions of years. These methods are discussed in Geologic Time.

Potassium-argon dating and uranium-series dating can be useful to the paleontologist, inasmuch as they aid in determining the age of the geologic samples in which the remains of life-forms are found. Nothing is simple, however, when it comes to dating specimens from the distant past. After all, geologists working in the realm of stratigraphy face numerous challenges in judging the age of samples, even with these sophisticated forms of radiometric dating. (For more on this subject, see Stratigraphy.)

In fact, the work of a paleontologist is much like that of the stratigrapher. The success of either type of scientist relies more on detailed and painstaking detective work than it does on sophisticated technology. Both must analyze lay-



STINGRAY FOSSIL, POSSIBLY DATING TO THE JURASSIC PERIOD. THIS IS A RARE FIND, BECAUSE STINGRAYS HAVE NO BONE, ONLY CARTILAGE, WHICH MAKES IT HARDER FOR THEM TO UNDERGO MINERALIZATION AND BE PRESERVED AS FOSSILS. *(© Gary Retherford/Photo Researchers. Reproduced by permission.)*

ered samples from the past to form a picture of the chronology in which the samples evolved, and oftentimes the apparent evidence can be deceptive. The principal difference between stratigraphic and paleontologic study relates to the materials: rocks in the first instance and fossils in the second.

REAL-LIFE APPLICATIONS

FOSSILS AND FOSSILIZATION

The term fossil refers to the remains of any prehistoric life-form, especially those preserved in rock before the end of the last ice age. The process by which a once-living thing becomes a fossil is known as fossilization. Generally, fossilization refers to changes in the hard portions, including bones, teeth, shells, and so on. This series of changes, in which minerals are replaced by different minerals, is known as mineralization. Sometimes, soft parts may experience mineralization and thus be preserved as fossils. A deceased organism in the process of becoming a fossil is known as a subfossil.

The majority of fossils come from invertebrates, such as mussels, that possess hard parts. Generally speaking, the older and smaller the organism, the more likely it is to have experienced fossilization, though other factors (which we will discuss later) also play a part. One of the most important factors involves location: for the most part, the lower the altitude, the greater the likelihood that a region will contain fossils. The best place of all is in the ocean, particularly the ocean floor. Nonetheless, fossils have been found on every continent of Earth, and the great distances that sometimes separate samples of the same species have aided earth scientists of many fields in understanding the processes that shaped our planet.

FOSSILS AND GEOLOGIC HISTORY. The earth beneath our feet is not standing still; rather, it is constantly moving, and over the great stretches of geologic time, the positions of the continents have shifted considerably. The details of these shifts are discussed in Plate Tectonics, an area of geologic study that explains much about the earth, from earthquakes and volcanoes to continental drift.

Paleontology has contributed to the study of plate tectonics by revealing apparent anomalies, such as fossilized dinosaur parts in Antarctica. No dinosaur could have lived on that forbidding continent, so there must be some other explanation: the continental plates themselves have moved. Long ago, the present continents were united in a "supercontinent" called Pangaea. When Pangaea split apart to form the present continents, the remains of various species were separated from one another and from the latitudes to which they were accustomed in life.

Fossilized remains of single-cell organisms have been found in rock samples as old as 3.5 Ga, and animal fossils have been located in rocks that date to the latter part of Precambrian time, as old as 1 Ga. Just as paleontologists have benefited from studies in chronostratigraphy and geochronometry, realms of stratigraphy concerned with the dating of rock samples, stratigraphers and other geologists have used fossil samples to date the rock strata in which they were found. Not all fossilized life-forms are equally suited to this purpose. Certain ones, known as index fossils or indicator species, have been associated strongly with particular intervals of geologic time. An example is the ammonoid, a mollusk that proliferated for about 350 Ma from the late Devonian to the early Cretaceous before experiencing mass extinction.

MAKING THE GRADE AS A FOSSIL. Everything that is living eventually dies, but not nearly all living things will become fossils. And even if they do, there are numerous reasons why fossils might not be preserved in such a way as to provide meaningful evidence for a paleontologist many millions of years later. In the potential pool of candidates for fossilization, as we have noted, organisms without hard structural portions are unlikely to become fossilized. Fossilization of soft-bodied creatures sometimes occurs, however, as, for instance, at Burgess Shale in British Columbia, where environmental conditions made possible the preservation of a wide range of samples.

Furthermore, location is a powerful factor. Sedimentary rock, formed by compression and deposition (i.e., formation of deposits) on the part of other rock and mineral particles, provides the setting for many fossils. Best of all is sediment, such as sand or mud, that has not yet consolidated into harder sandstone, limestone, or other rocks. Organisms that die in upland locations are more likely to be disturbed either by wind or by scavengers, creatures that feed on the remains of living things. On the other hand, an organism at the bottom of an ocean is out of reach from most scavengers. Even at lesser depths, if the organism is in a calm, relatively scavenger-free marine environment, there is a good chance that it will be preserved.

Assuming that all the conditions are right and the dead organism is capable of undergoing fossilization, it will experience mineralization of one type or another. Living things already contain minerals, which is the reason why people take mineral supplements to augment the substances nature has placed in their bodies to preserve and extend life. In the mineralization of a fossil, the minerals in the organism's body may be replaced by other ones, or other minerals may be added to existing ones. It is also possible that both the hard and soft parts will dissolve and be replaced by a mineral cement that forms a mold that preserves the shape of the organism.

FINDING AND STUDYING FOSSILS. Only about 30% of species are ever fossilized, a fact that scientists must take into account, because it could skew their reading of the paleontologic record. If a paleontologist judges the past only from the fossils that have been found in an area, it will result in a picture of a past environment that contained only certain species, when, in fact, others were present. Furthermore, there are many factors that contribute to the loss of fossils. For instance, if the area has been subjected to violent tectonic activity, it is likely that the sample will be destroyed partially or wholly.

The removal of a fossil from its home in the rock is a painstaking process akin to restoring a valuable piece of art. Before removing it, the paleontologist photographs the fossil and surrounding strata and records details about the environment. Only when these steps have been taken is the fossil removed. This is done with a rock saw, which is used to cut out carefully a large area surrounding the fossil. The sample is then jacketed, or wrapped in muslin with an additional layer of wet plaster, and taken to a laboratory for study.

Fossil research can reveal a great deal about the history of life on Earth, including the relationships between species or between species and their habitats. Studies of dinosaur bones have brought to light proteins that existed in the bodies of these long-gone creatures, while research on certain oxygen isotopes has aided attempts to discover whether dinosaurs were warm-blooded creatures. Thanks to advances in the understanding of DNA (deoxyribonucleic acid), which provides the genetic codes for all living things, it may be possible to make even more detailed studies in the future.

Mass Extinction

The remains of dinosaurs, of course, have an importance aside from their significance to paleontology. The bodies of these giant lizards have been deposited in the earth, where over time they became coal, peat, petroleum, and other fossil fuels. The latter are discussed in Economic Geology, but the fact that the dinosaurs disappeared at all is of particular interest to paleontology. Why are there no dinosaurs roaming the earth today? The answer appears to be that they were wiped out in a dramatic event, perhaps brought about as the result of a meteorite impact.

Numerous species have become extinct, typically as a result of their inability to adapt to changes in their natural environment. More recently, some extinctions or endangerments of species have been attributed to human activities, including hunting and the disruption of natural habitats. For the most part, however, extinction is simply a part of Earth's history, a result of the fact that nature has a way of destroying organisms that do not adapt (the "survival of the fittest"). But there have been occasions in the course of the planet's past in which vast numbers of individuals and species perished at once. A natural catastrophe may destroy a large population of individuals within a locality, a phenomenon known as mass mortality. Or mass mortality may take place on a global scale, destroying many species, in which case it is known as mass extinction.

CAUSES OF MASS EXTINCTION. The Bible depicts an example of near mass extinction, in the form of Noah's flood, and, indeed, several instances of mass extinction have resulted from sudden and dramatic changes in ocean levels. Others have been caused by tectonic events, most notably vast volcanic eruptions that filled the atmosphere with so much dust that they caused a violent change in temperature. Scientific speculation concerning other such extinctions has pointed to events in or from space—either the explosion of a star or the impact of a meteorite on Earth—as the cause of atmospheric changes and hence mass extinction.

Even though scientists have a reasonable idea of the immediate causes of mass extinction in some cases, their understanding of the ultimate or root causes is still limited. This fact was expressed by the University of Chicago paleobiologist David M. Raup, who wrote: "The disturbing reality is that for none of the thousands of well-documented extinctions in the geologic past do we have a solid explanation of why the extinction occurred."

EXAMPLES OF MASS EXTINCTION. The five largest known mass extinctions occurred at intervals of 50 Ma to 100 Ma over a span of time from about 435 to 65 million years ago. Most occurred at the end of a period, which is no accident, since geologists have used mass extinction as a factor in determining the parameters of a specific period.

In the late Ordovician period, about 435 Ma ago, a drop in the ocean level wiped out one-fourth of all marine families. Similarly, changes in sea level, along with climate changes, appear to have caused the destruction of one-fifth of existing marine families during the late Devonian period (about 357 Ma ago). Worst of all was the "great dying," as the extinction at the end of the Permian period (about 250 Ma) is known. Perhaps caused by a volcanic eruption in Siberia, it eliminated a staggering 96% of all species over a period of about a million years.

During the late Triassic period, about 198 million years ago, another catastrophe eliminated a quarter of marine families. Paleontologists know this, as they know about other mass extinctions, by the inordinate numbers of fossilized samples found in rock strata dating to that period. This reliance on the fossil record is also reflected in the fact that the scope of early mass extinctions usually is expressed in terms of marine life. As we have seen, the ocean environment provides the most reliable fossil record. Creatures died on land as well, but the terrestrial record is simply less reliable or less complete.

Scientific disagreement over the late-Triassic mass extinction exemplifies the fact that our knowledge of these distant events is not firmly established, but rather is subject to much scientific conjecture and dispute. (This does not mean that just any old idea can compete on an equal

KEY TERMS

ABSOLUTE AGE: The absolute age of a geologic phenomenon is its age in Earth years. Compare with *relative age.*

ATMOSPHERE: In general, an atmosphere is a blanket of gases surrounding a planet. Unless otherwise identified, however, the term refers to the atmosphere of Earth, which consists of nitrogen (78%), oxygen (21%), argon (0.93%), and other substances that include water vapor, carbon dioxide, ozone, and noble gases such as neon, which together comprise 0.07%.

BIOSPHERE: A combination of all living things on Earth—plants, mammals, birds, reptiles, amphibians, aquatic life, insects, viruses, single-cell organisms, and so on—as well as all formerly living things that have not yet decomposed.

BIOSTRATIGRAPHY: An area of stratigraphy involving the study of fossilized plants and animals to establish dates for and correlations between stratigraphic layers.

CHRONOSTRATIGRAPHY: A subdiscipline of stratigraphy devoted to studying the relative ages of rocks. Compare with *geochronometry.*

CONTINENTAL DRIFT: The theory that the configuration of Earth's continents was once different than it is today; that some of the individual landmasses of today once were joined in other continental forms; and that these landmasses later separated and moved to their present locations.

CORRELATION: A method of establishing age relationships between various rock strata. There are two basic types of correlation: physical correlation, which requires comparison of the physical characteristics of the strata, and fossil correlation, the comparison of fossil types.

DATING: Any effort directed toward finding the age of a particular item or phenomenon. Methods of geologic dating are either relative (i.e., comparative and usually based on rock strata) or absolute. The latter, based on such methods as the study of radioactive isotopes, typically is given in terms of actual years or millions of years.

EON: The longest phase of geologic time. Earth's history has consisted of four eons, the Hadean or Priscoan, Archaean, Proterozoic, and Phanerozoic. The next-smallest subdivision of geologic time is the era.

EPOCH: The fourth-longest phase of geologic time, shorter than an era and longer than an age or a chron. The current epoch is the Holocene, which began about 0.01 Ma (10,000 years) ago.

ERA: The second-longest phase of geologic time, after an eon. The current eon, the Phanerozoic, has had three eras, the Paleozoic, Mesozoic, and Cenozoic, which is the current era. The next-smallest subdivision of geologic time is the period.

FOSSIL: The mineralized remains of any prehistoric life-form, especially those preserved in rock before the end of the last ice age.

FOSSILIZATION: The process by which a once-living organism becomes a fossil. Generally, fossilization involves mineralization of the organism's hard portions, such as bones, teeth, and shells.

KEY TERMS CONTINUED

GA: An abbreviation meaning "gigayears," or "billion years." The age of Earth is about 4.6 Ga.

GEOCHRONOMETRY: An area of stratigraphy devoted to determining absolute dates and time intervals. Compare with *chronostratigraphy.*

GEOLOGIC TIME: The vast stretch of time over which Earth's geologic development has occurred. This span (about 4.6 billion years) dwarfs the history of human existence, which is only about two million years. Much smaller still is the span of human civilization, only about 5,500 years.

HISTORICAL GEOLOGY: The study of Earth's physical history. Historical geology is one of two principal branches of geology, the other being physical geology.

INVERTEBRATE: An animal without an internal skeleton.

ISOTOPES: Atoms that have an equal number of protons, and hence are of the same element, but differ in their number of neutrons. This results in a difference of mass. An isotope may be either stable or radioactive.

LAW OF FAUNAL SUCCESSION: The principle that all samples of any given fossil species were deposited on Earth, regardless of location, at more or less the same time. This makes it possible to correlate widely separated strata.

MA: An abbreviation used by earth scientists, meaning "million years," or "megayears." When an event is designated as, for instance, 160 Ma, it usually means 160 million years ago.

MASS EXTINCTION: A phenomenon in which numerous species cease to exist at or around the same time, usually as the result of a natural calamity.

MINERALIZATION: A series of changes experienced by a once-living organism during fossilization. In mineralization, minerals in the organism are either replaced or augmented by different minerals, or the hard portions of the organism dissolve completely.

ORGANIC: At one time, chemists used the term *organic* only in reference to living things. Now the word is applied to most compounds containing carbon, with the exception of carbonates (which are minerals), and oxides, such as carbon dioxide.

PALEOBOTANY: An area of paleontology involving the study of past plant life.

PALEOECOLOGY: An area of paleontology devoted to studying the relationship between prehistoric plants and animals and their environments.

PALEONTOLOGY: The study of lifeforms from the distant past, primarily as revealed through the fossilized remains of plants and animals.

PALEOZOOLOGY: An area of paleontology devoted to the study of prehistoric animal life.

PERIOD: The third-longest phase of geologic time, after an era. The current eon, the Phanerozoic, has had 11 periods, and the current era, the Cenozoic, has consisted of three periods, of which the most recent is the Quaternary. The next-smallest subdivision of geologic time is the epoch.

KEY TERMS CONTINUED

PRECAMBRIAN TIME: A term that refers to the first three of four eons in Earth's history, which lasted from about 4,560 to about 545 Ma ago.

RADIOACTIVITY: A term describing a phenomenon whereby certain materials are subject to a form of decay brought about by the emission of high-energy particles or radiation. Forms of particles or energy include alpha particles (positively charged helium nuclei), beta particles (either electrons or subatomic particles called positrons, or gamma rays, which occupy the highest energy level in the electromagnetic spectrum.

RADIOMETRIC DATING: A method of absolute dating using ratios between "parent" isotopes and "daughter" isotopes, which are formed by the radioactive decay of parent isotopes. Radiometric dating also may involve ratios between radioactive isotopes and stable isotopes.

RELATIVE AGE: The relative age of a geologic phenomenon is its age compared with other geologic phenomena, particularly the stratigraphic record of rock layers. Compare with *absolute age.*

SEDIMENT: Material deposited at or near Earth's surface from a number of sources, most notably preexisting rock.

SEDIMENTARY ROCK: Rock formed by compression and deposition (i.e., formation of deposits) on the part of other rock and mineral particles. Sedimentary rock is one of the three major types of rock, along with igneous and metamorphic.

SEDIMENTOLOGY: The study and interpretation of sediments, including sedimentary processes and formations.

STRATA: Layers, or beds, of rocks beneath Earth's surface. The singular form is *stratum.*

STRATIGRAPHY: The study of rock layers, or strata, beneath Earth's surface.

VERTEBRATE: An animal with an internal skeleton.

footing: we are talking here about differences of opinion among highly trained specialists.) At any rate, some scientists refer to the late-Triassic mass extinction as being one of the less exciting or eventful mass extinctions. Of course, it is hard to see how a mass extinction could be unexciting or uneventful, but they mean this in comparative terms; on the other hand, some paleontologists maintain that the late-Triassic was among the most devastating.

As to the cause, some theorists point to a group of impact sites spread across Canada, the northern United States, and Ukraine, places that would have been more or less contiguous at the time of the mass extinction. Difficulties in analyzing the "signatures" left by the projectiles that made these impressions have prevented theorists from saying with any degree of certainty whether it was a comet or an asteroid that caused the impact. Others, in particular a team from the University of California at Berkeley led by geologist Paul R. Renne, cite a volcanic eruption as either the cause of the mass extinction, or at least a major abetting factor to an extinction already in progress. According to Renne and his team, basalt outcroppings scattered from New Jersey to Brazil to west Africa (again, areas that would have been contiguous then) suggest that a volcanic eruption of almost inconceivable magnitude occurred about 200 million years ago. Such an eruption would surely have destroyed vast quantities of living things.

The last and best known mass extinction occurred about 65 million years ago, marking the end of the Cretaceous period—and the end of the dinosaurs. As to what happened, paleontologists and other scientists have proposed a number of theories: a rapid climate change; the emergence of new poisonous botanical species, eaten by herbivorous dinosaurs, that resulted in the passing of toxins along the food web (see Ecosystems); an inability to compete successfully with the rapidly evolving mammals; and even an epidemic disease to which the dinosaurs possessed no immunity.

Interesting as many of these theories are, none has gained anything like the widespread acceptance achieved by another scenario. According to this highly credible theory, an asteroid hit Earth, hurtling vast quantities of debris into the atmosphere, blocking out the sunlight, and greatly lowering Earth's surface temperature. Around the world, geologists have found traces of iridium deposited at a layer equivalent to the boundary between the Cretaceous and Tertiary periods, the Tertiary being the beginning of the present Cenozoic era. This is significant, because iridium seldom appears on Earth's surface—but it *is* found in asteroids.

HUMANS AND MASS EXTINCTION. There have been much more recent, if less dramatic, examples of mass extinction, including those caused by the most highly developed of all life-forms: humans. Among these examples are the well-documented (and *very* recent) mass extinctions brought on by destruction of tropical rainforests. Such activities are killing off a vast array of organisms: according to the highly respected Harvard biologist Edward O. Wilson, some 17,500 species are disappearing each year. But cases of mass extinction are not limited to modern times.

When prehistoric hunters (the ancestors of today's Native Americans) crossed the Bering land bridge from Siberia to Alaska some 12,000 years ago, they found an array of species unknown in the Americas today. These species included mammoths and mastodons; giant bears, beaver, and bison; and even saber-toothed tigers, camels, and lions. Perhaps most remarkable of all, it appears that prehistoric America was once home to a creature that would prove to be of enormous benefit to humans until the beginning of the automotive age: the horse. Horses did not reappear in the Americas until Europeans arrived to conquer those lands after A.D. 1500.

Were Dinosaurs Warm-Blooded?

One of the most significant scientific debates of the later twentieth and early twenty-first centuries, not only in paleontology but in the earth sciences or even science itself, is the question of whether or not the dinosaurs were warm-blooded. In other words, were they like modern reptiles, which must adjust their temperature by moving into the sunlight when they are cold, and into the shade when they are too hot? Or were they more like modern birds and mammals, whose bodies generate their own heat?

A warm-blooded animal always has a more or less constant body temperature, regardless of the temperature of its environment. This is due to the fact that it produces heat by the burning of food, as well as by physical activity, and stores that heat under a layer of fat just beneath the skin. Warm-blooded animals are also capable of cooling down their bodies by perspiring and panting. Birds and mammals are the only warm-blooded animals; all others are cold-blooded. A cold-blooded creature, on the other hand, lacks control over its body temperature and therefore is warm when its environment is warm, and cold when its environment is cold.

The difference between warm- and cold-blooded animals is partly one of metabolic rate, or the rate at which nutrients are broken down and converted into energy. Cold-blooded creatures have slow metabolic rates; think of a python that swallows a medium-sized mammal whole and takes several days to digest it. The dinosaur debate is therefore often framed as a question of whether the dinosaurs' bodies had a relatively high or relatively low metabolic rate.

THE DEBATE. Until the 1960s, there was no debate: dinosaurs, whose existence had been known for about a century, were assumed to be big, dumb, slow, cold-blooded creatures. Then, in 1968, Robert T. Bakker—an undergraduate at Yale University, not a professor or a full-fledged paleontologist—revolutionized the world of paleontology with a paper called "The Superiority of Dinosaurs."

In his article, Bakker described dinosaurs as "fast, agile, energetic creatures" whose physiology was so advanced that even the biggest and heaviest of them could outrun a human. Just a year later, John H. Ostrom, a professor of paleontology who also happened to be at Yale, wrote that a recently identified species of theropod dinosaur must have been "an active and very agile predator."

Thus began the great dinosaur debate, which rages even today. *Jurassic Park* reinforced the Bakker-Ostrom position, portraying *Velociraptor* as a cunning, fast-moving predator with clear links to birds. And indeed there are many arguments for endothermy (warm-bloodedness) in dinosaurs—arguments that relate to everything from brain size to rate of growth to the latitudes at which dinosaur fossils have been located. On the other hand, there is plenty of evidence for ectothermy (cold-bloodedness), based on the dinosaurs' size, scaliness, the climate in the Mesozoic era, and so on.

To explore, compare, and judge these many arguments, the reader is encouraged to consult the "Were Dinosaurs Warm-Blooded?" Web site listed in the "Where to Learn More" section at the conclusion of this essay. However, a word of warning, as noted on that site: "The issue is a tangled, complex one. There are not just two sides to the issue; there are numerous competing hypotheses. If you're looking for a major controversy in science, look no further!"

WHERE TO LEARN MORE

Cadbury, Deborah. *Terrible Lizard: The First Dinosaur Hunters and the Birth of a New Science.* New York: Holt, 2001.

K-12: Paleontology: Dinos (Web site). <http://www.ceismc.gatech.edu/busyt/paleo.html>.

Morris, S. Conway. *The Crucible of Creation: The Burgess Shale and the Rise of Animals.* New York: Oxford University Press, 1998.

Munro, Margaret, and Karen Reczuch. *The Story of Life on Earth.* Toronto: Douglas and McIntyre, 2000.

Oceans of Kansas Paleontology (Web site). <http://www.oceansofkansas.com/>.

Paleontology and Fossils Resources—University of Arizona Library, Tucson (Web site). <http://www.library.arizona.edu/users/mount/paleont.html>.

Palmer, Douglas. *Atlas of the Prehistoric World.* Bethesda, MD: Discovery Communications, 1999.

Sanders, Robert. "New Evidence Links Mass Extinction with Massive Eruptions That Split Pangea Supercontinent and Created the Atlantic 200 Million Years Ago." University of California, Berkeley. (Web site). http://www.berkeley.edu/news/media/releases/99legacy/4-22-1999b.html>.

Singer, Ronald. *Encyclopedia of Paleontology.* Chicago: Fitzroy Dearborn Publishers, 1999.

University of California, Berkeley Museum of Paleontology (Web site). <http://www.ucmp.berkeley.edu/>.

USGS (United States Geological Survey) Paleontology Home Page (Web site). <http://geology.er.usgs.gov/paleo/>.

"Were Dinosaurs Warm-Blooded?" (Web site). <http://pubs.usgs.gov/gip/dinosaurs/warmblood.html>.

MINERALS

CONCEPT

A mineral is a naturally occurring, typically inorganic substance with a specific chemical composition and structure. An unknown mineral usually can be identified according to known characteristics of specific minerals in terms of certain parameters that include its appearance, its hardness, and the ways it breaks apart when fractured. Minerals are not to be confused with rocks, which are typically aggregates of minerals. There are some 3,700 varieties of mineral, a handful of which are abundant and wide-ranging in their application. Many more occur less frequently but are extremely important within a more limited field of uses.

HOW IT WORKS

Introduction to Minerals

The particulars of the mineral definition deserve some expansion, especially inasmuch as mineral has an everyday definition somewhat broader than its scientific definition. In everyday usage, minerals would be the natural, nonliving materials that make up rocks and are mined from the earth. According to this definition, minerals would include all metals, gemstones, clays, and ores. The scientific definition, on the other hand, is much narrower, as we shall see.

The fact that a mineral must be inorganic brings up another term that has a broader meaning in everyday life than in the world of science. At one time, the scientific definition of organic was more or less like the meaning assigned to it by nonscientists today, as describing all living or formerly living things, their parts, and substances that come from them. Today, however, chemists use the word organic to refer to any compound that contains carbon bonded to hydrogen, thus excluding carbonates (which are a type of mineral) and oxides such as carbon dioxide or carbon monoxide. Because a mineral must be inorganic, this definition eliminates coal and peat, both of which come from a wide-ranging group of organic substances known as hydrocarbons.

A mineral also occurs naturally, meaning that even though there are artificial substances that might be described as "mineral-like," they are not minerals. In this sense, the definition of a mineral is even more restricted than that of an element, discussed later in this essay, even though there are nearly 4,000 minerals and more than 92 elements. The number 92, of course, is not arbitrary: that is the number of elements that occur in nature. But there are additional elements, numbering 20 at the end of the twentieth century, that have been created artificially.

PHYSICAL AND CHEMICAL PROPERTIES OF MINERALS. The specific characteristics of minerals can be discussed both in physical and in chemical terms. From the standpoint of physics, which is concerned with matter, energy, and the interactions between the two, minerals would be described as crystalline solids. The definition of a mineral is narrowed further in terms of its chemistry, or its atomic characteristics, since a mineral must be of unvarying composition.

A mineral, then, must be solid under ordinary conditions of pressure and temperature. This excludes petroleum, for instance (which, in any case, would have been disqualified owing to its organic origins), as well as all other liquids and gases. Moreover, a mineral cannot be just any type of solid but must be a crystalline one—that is, a solid in which the constituent parts have a simple and definite geometric arrangement that is repeated in all directions. This rule, for instance, eliminates clay, an example of an amorphous solid.

Chemically, a mineral must be of unvarying composition, a stipulation that effectively limits minerals to elements and compounds. Neither sand nor glass, for instance, is a mineral, because the composition of both can vary. Another way of putting this is to say that all minerals must have a definite chemical formula, which is not true of sand, dirt, glass, or any other mixture. Let us now look a bit more deeply into the nature of elements and compounds, which are collectively known as pure substances, so as to understand the minerals that are a subset of this larger grouping.

Elements

The periodic table of elements is a chart that appears in most classrooms where any of the physical sciences are taught. It lists all elements in order of atomic number, or the number of protons (positively charged subatomic particles) in the atomic nucleus. The highest atomic number of any naturally occurring element is 92, for uranium, though it should be noted that a very few elements with an atomic number lower than 92 have never actually been found on Earth. On the other hand, *all* elements with an atomic number higher than 92 are artificial, created either in laboratories or as the result of atomic testing.

An element is a substance made of only one type of atom, meaning that it cannot be broken down chemically to create a simpler substance. In the sense that each is a fundamental building block in the chemistry of the universe, all elements are, as it were, "created equal." They are not equal, however, in terms of their abundance. The first two elements on the periodic table, hydrogen and helium, represent 99.9% of the matter in the entire universe. Though Earth contains little of either, our planet is only a tiny dot within the vastness of space; by contrast, stars such as our Sun are composed almost entirely of those elements (see Sun, Moon, and Earth).

ABUNDANCE ON EARTH. Of all elements, oxygen is by far the most plentiful on Earth, representing nearly half—49.2%—of the total mass of atoms found on this planet. (Here the term mass refers to the known elemental mass of the planet's atmosphere, waters, and crust; below the crust, scientists can only speculate, though it is likely that much of Earth's interior consists of iron.)

Together with silicon (25.7%), oxygen accounts for almost exactly three-fourths of the elemental mass of Earth. If we add in aluminum (7.5%), iron (4.71%), calcium (3.39%), sodium (2.63%), potassium (2.4%), and magnesium (1.93%), these eight elements make up about 97.46% of Earth's material. Hydrogen, so plentiful in the universe at large, ranks ninth on Earth, accounting for only 0.87% of the planet's known elemental mass. Nine other elements account for a total of 2% of Earth's composition: titanium (0.58%), chlorine (0.19%), phosphorus (0.11%), manganese (0.09%), carbon (0.08%), sulfur (0.06%), barium (0.04%), nitrogen (0.03%), and fluorine (0.03%). The remaining 0.49% is made up of various other elements.

Looking only at Earth's crust, the numbers change somewhat, especially at the lower end of the list. Listed below are the 12 most abundant elements in the planet's crust, known to earth scientists simply as "the abundant elements." These 12, which make up 99.23% of the known crustal mass, together form approximately 40 different minerals that account for the vast majority of that 99.23%. Following the name and chemical symbol of each element is the percentage of the crustal mass it composes.

Abundance of Elements in Earth's Crust

- Oxygen (O): 45.2%
- Silicon (Si): 27.2%
- Aluminum (Al): 8.0%
- Iron (Fe): 5.8%
- Calcium (Ca): 5.06%
- Magnesium (Mg): 2.77%
- Sodium (Na): 2.32%
- Potassium (K): 1.68%
- Titanium (Ti): 0.86%
- Hydrogen (H): 0.14%
- Manganese (Mn): 0.1%
- Phosphorus (P): 0.1%

Atoms, Molecules, and Bonding

As noted earlier, an element is identified by the number of protons in its nucleus, such that any atom with six protons *must* be carbon, since carbon has an atomic number of 6. The number of electrons, or negatively charged subatomic particles, is the same as the number of protons, giving an atom no net electric charge.

An atom may lose or gain electrons, however, in which case it becomes an ion, an atom or group of atoms with a net electric charge. An atom that has gained electrons, and thus has a negative charge, is called an anion. On the other hand, an atom that has lost electrons, thus becoming positive in charge, is a cation.

In addition to protons and electrons, an atom has neutrons, or neutrally charged particles, in its nucleus. Neutrons have a mass close to that of a proton, which is much larger than that of an electron, and thus the number of neutrons in an atom has a significant effect on its mass. Atoms that have the same number of protons (and therefore are of the same element), but differ in their number of neutrons, are called isotopes.

COMPOUNDS AND MIXTURES. Whereas there are only a very few elements, there are millions of compounds, or substances made of more than one atom. A simple example is water, formed by the bonding of two hydrogen atoms with one oxygen atom; hence the chemical formula for water, which is H_2O. Note that this is quite different from a mere mixture of hydrogen and oxygen, which would be something else entirely. Given the gaseous composition of the two elements, combined with the fact that both are extremely flammable, the result could hardly be more different from liquid water, which, of course, is used for putting out fires.

The difference between water and the hydrogen-oxygen mixture described is that whereas the latter is the result of mere physical mixing, water is created by chemical bonding. Chemical bonding is the joining, through electromagnetic attraction, of two or more atoms to create a compound. Of the three principal subatomic particles, only electrons are involved in chemical bonding—and only a small portion of those, known as valence electrons, which occupy the outer shell of an atom. Each element has a characteristic pattern of valence electrons, which determines the ways in which the atom bonds.

CHEMICAL BONDING. Noble gases, of which helium is an example, are noted for their lack of chemical reactivity, or their resistance to bonding. While studying these elements, the German chemist Richard Abegg (1869–1910) discovered that they all have eight valence electrons. His observation led to one of the most important principles of chemical bonding: atoms bond in such a way that they achieve the electron configuration of a noble gas. This concept, known as the octet rule, has been shown to be the case in most stable chemical compounds.

Abegg hypothesized that atoms combine with one another because they exchange electrons in such a way that both end up with eight valence electrons. This was an early model of ionic bonding, which results from attractions between ions with opposite electric charges: when they bond, these ions "complete" each other. Metals tend to form cations and bond with nonmetals that have formed anions. The bond between anions and cations is known as an ionic bond, and is extremely strong.

The other principal type of bond is a covalent bond. The result, once again, is eight valence electrons for each atom, but in this case, the nuclei of the two atoms share electrons. Neither atom "owns" them; rather, they share electrons. Today, chemists understand that most bonds are neither purely ionic nor purely covalent; instead, there is a wide range of hybrids between the two extremes, which are a function of the respective elements' electronegativity, or the relative ability of an atom to attract valence electrons. If one element has a much higher electronegativity value than the other one, the bond will be purely ionic, but if two elements have equal electronegativity values, the bond is purely covalent. Most bonds, however, fall somewhere between these two extremes.

INTERMOLECULAR BONDING. Chemical bonds exist between atoms and within a molecule. But there are also bonds *between* molecules, which affect the physical composition of a substance. The strength of intermolecular bonds is affected by the characteristics of the interatomic, or chemical, bond.

For example, the difference in electronegativity values between hydrogen and oxygen is great enough that the bond between them is not purely covalent, but instead is described as a polar covalent bond. Oxygen has a much higher electronegativity (3.5) than hydrogen (2.1), and therefore the electrons tend to gravitate toward the oxygen atom. As a result, water molecules have a strong negative charge on the side occupied by the oxygen atom, with a resulting positive charge on the hydrogen side.

By contrast, molecules of petroleum, a combination of carbon and hydrogen, tend to be nonpolar, because carbon (with an electronegativity value of 2.5) and hydrogen have very similar electronegativity values. Therefore the electric charges are more or less evenly distributed in the molecule. As a result, water molecules form strong attractions, known as dipole-dipole attractions, to each other. Molecules of petroleum, on the other hand, have little attraction to each other, and the differences in charge distribution account for the fact that water and oil do not mix.

Even weaker than the bonds between nonpolar molecules, however, are those between highly reactive elements, such as the noble gases and the "noble metals"—gold, silver, and copper, which resist bonding with other elements. The type of intermolecular attraction that exists in such a situation is described by the term London dispersion forces, a reference to the German-born American physicist Fritz Wolfgang London (1900–1954).

The bonding between molecules of most other metals, however, is described by the electron sea model, which depicts metal atoms as floating in a "sea" of valence electrons. These valence electrons are highly mobile within the crystalline structure of the metal, and this mobility helps explain metals' high electric conductivity. The ease with which metal crystals allow themselves to be rearranged explains not only metals' ductility (their ability to be shaped) but also their ability to form alloys, a mixture containing two or more metals.

The Crystalline Structure of Minerals

By definition, a solid is a type of matter whose particles resist attempts at compression. Because of their close proximity, solid particles are fixed in an orderly and definite pattern. Within the larger category of solids are crystalline solids, or those in which the constituent parts are arranged in a simple, definite geometric pattern that is repeated in all directions.

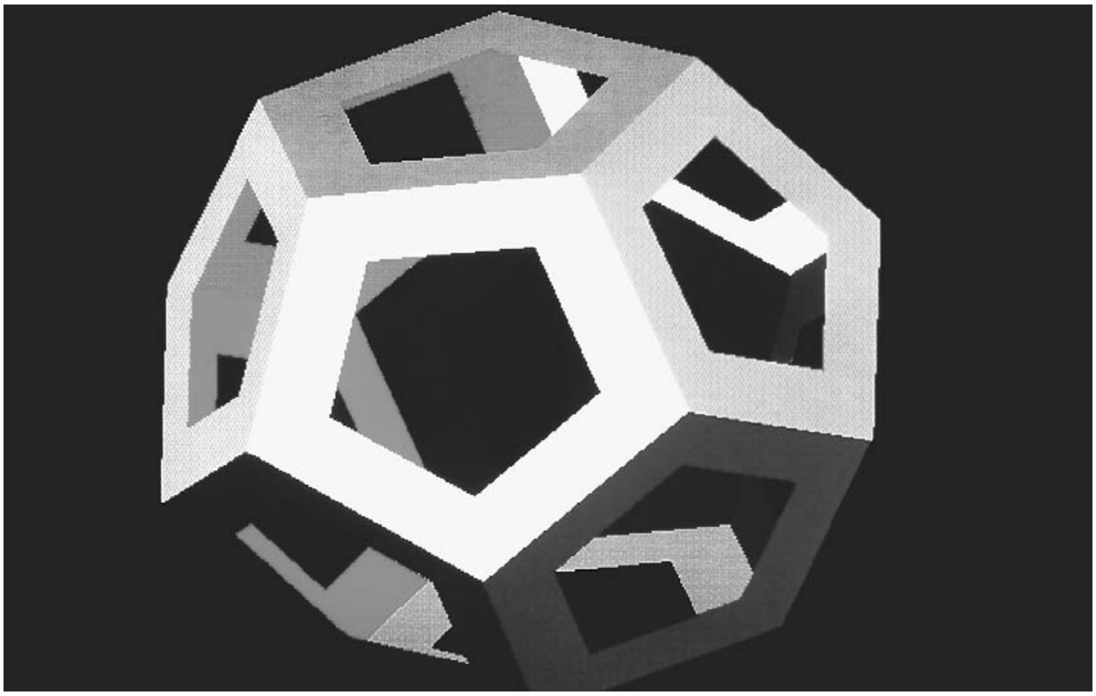
The term crystal is popularly associated with glass and with quartz, but only one of these is a crystalline solid. Quartz is a member of the silicates, a large group of minerals that we will discuss later in this essay. Glass, on the other hand, is an amorphous solid, meaning that its molecules are not arranged in an orderly pattern.

CRYSTAL SYSTEMS. Elsewhere in this book (Earth, Science, and Nonscience and Planetary Science), there is considerable discussion of misconceptions originating with Aristotle (384–322 B.C.). Despite his many achievements, including significant contributions to the biological sciences, the great Greek philosopher spawned a number of erroneous concepts, which prevailed in the physical sciences until the dawn of the modern era. At least Aristotle made an attempt at scientific study, however; for instance, he dissected dead animals to observe their anatomic structures. His teacher, Plato (427?–347 B.C.), on the other hand, is hardly ever placed among the ranks of those who contributed, even ever so slightly, to progress in the sciences.

There is a reason for this. Plato, in contrast to his pupil, made virtually no attempt to draw his ideas about the universe from an actual study of it. Within Plato's worldview, the specific qualities of any item, including those in the physical world, reflected the existence of perfect and pure ideas that were more "real" than the physical objects themselves. Typical of his philosophy was his idea of the five Platonic solids, or "perfect" geometric shapes that, he claimed, formed the atomic substructure of the world.

The "perfection" of the Platonic solids lay in the fact that they are the only five three-dimensional objects in which the faces constitute a single type of polygon (a closed shape with three or more sides, all straight), while the vertices (edges) are all alike. These five are the tetrahedron, octahedron, and icosahedron, composed of equilateral triangles (four, eight, and twenty, respectively); the cube, which, of course, is made of six squares; and the dodecahedron, made up of twelve pentagons. Plato associated the latter solid with the shape of atoms in outer space, while the other four corresponded to what the Greeks believed were the elements on Earth: fire (tetrahedron), earth (cube), air (octahedron), and water (icosahedron).

All of this, of course, is nonsense from the standpoint of science, though the Platonic solids are of interest within the realm of mathematics. Yet amazingly, Plato in his unscientific way actually touched on something close to the truth, as applied to the crystalline structure of minerals.



A dodecahedron, one of the Platonic solids. *(© Richard Duncan/Photo Researchers. Reproduced by permission.)*

Despite the large number of minerals, there are just six crystal systems, or geometric shapes formed by crystals. For any given mineral, it is possible for a crystallographer (a type of mineralogist concerned with the study of crystal structures) to identify its crystal system by studying a good, well-formed specimen, observing the faces of the crystal and the angles at which they meet.

An isometric crystal system is the most symmetrical of all, with faces and angles that are most clearly uniform. Because of differing types of polygon that make up the faces, as well as differing numbers of vertices, these crystals appear in 15 forms, several of which are almost eerily reminiscent of Plato's solids: not just the cube (exemplified by halite crystals) but also the octahedron (typical of spinels) and even the dodecahedron (garnets).

REAL-LIFE APPLICATIONS

Mineral Groups

Before the time of the great German mineralogist Georgius Agricola (1494–1555), attempts to classify minerals were almost entirely overshadowed by the mysticism of alchemy, by other nonscientific preoccupations, or by simple lack of knowledge. Agricola's *De re metallica* (On minerals, 1556), published after his death, constituted the first attempt at scientific mineralogy and mineral classification, but it would be two and a half centuries before the Swedish chemist Jöns Berzelius (1779–1848) developed the basics of the classification system used today.

Berzelius's classification system was refined later in the nineteenth century by the American mineralogist James Dwight Dana (1813–1895) and simplified by the American geologists Brian Mason (1917–) and L. G. Berry (1914–). In general terms, the classification system accepted by mineralogists today is as follows:

- Class 1: Native elements
- Class 2: Sulfides
- Class 3: Oxides and hydroxides
- Class 4: Halides
- Class 5: Carbonates, nitrates, borates, iodates
- Class 6: Sulfates, chromates, molybdates, tungstates
- Class 7: Phosphates, arsenates, vanadates
- Class 8: Silicates

NATIVE ELEMENTS. The first group, native elements, includes (among other things) metallic elements that appear in pure form somewhere on Earth: aluminum, cadmium, chromium, copper, gold, indium, iron, lead, mercury, nickel, platinum, silver, tellurium, tin, titanium, and zinc. This may seem like a great number of elements, but it is only a small portion of the 87 metallic elements listed on the periodic table.

The native elements also include certain metallic alloys, a fact that might seem strange for several reasons. First of all, an alloy is a mixture, not a compound, and, second, people tend to think of alloys as being man-made, not natural. The list of metallic alloys included among the native elements, however, is very small, and they meet certain very specific mineralogic criteria regarding consistency of composition.

The native elements class also includes native nonmetals such as carbon, in the form of graphite or its considerably more valuable alter ego, diamond, as well as elemental silicon (an extremely important building block for minerals, as we shall see) and sulfur. For a full list of native elements and an explanation of criteria for inclusion, as well as similar data for the other classes of mineral, the reader is encouraged to consult the *Minerals by Name* Web site, the address of which is provided in "Where to Learn More" at the end of this essay.

SULFIDES AND HALIDES. Most important ores (a rock or mineral possessing economic value)—copper, lead, and silver—belong to the sulfides class, as does a mineral that often has been mistaken for a precious metal—iron sulfide, or pyrite. Better known by the colloquial term fool's gold, pyrite has proved valuable primarily to con artists who passed it off as the genuine article. During World War II, however, pyrite deposits near Ducktown, Tennessee, became valuable owing to the content of sulfur, which was extracted for use in defense applications.

Whereas the sulfides fit the common notion of a mineral as a hard substance, halides, which are typically soft and transparent, do not. Yet they are indeed a class of minerals, and they include one of the best-known minerals on Earth: halite, known chemically as NaCl or sodium chloride—or, in everyday language, table salt.

OXIDES. Oxides, as their name suggests, are minerals containing oxygen; however, if all oxygen-containing minerals were lumped into just one group, that group would take up almost the entire list. For instance, under the present system, silicates account for the vast majority of minerals, but since *those* contain oxygen as well, a list that grouped all oxygen-based minerals together would consist of only four classes: native elements, sulfides, halides, and a swollen oxide category that would include 90% of all known minerals.

Instead, the oxides class is limited only to noncomplex minerals that contain either oxygen or hydroxide (OH). Examples of oxides include magnetite (iron oxide) and corundum (aluminum oxide.) It should be pointed out that a single chemical name, such as iron oxide or aluminum oxide, is not limited to a single mineral; for example, anatase and brookite are both titanium oxide, but they represent different combinations.

OTHER NONSILICATES. All the mineral classes discussed to this point, as well as several others to follow, are called nonsilicates, a term that stresses the importance of silicates among mineral classes.

Like the oxides, the carbonates, or carbon-based minerals, are a varied group. This class also contains a large number of minerals, making it the most extensive group aside from silicates and phosphates. Among these are limestones and dolostones, some of the most abundant rocks on Earth.

The phosphates, despite their name, may or may not include phosphorus; in some cases, arsenic, vanadium, or antimony may appear in its place. The same is true of the sulfates, which may or may not involve sulfur; some include chromium, tungsten, selenium, tellurium, or molybdenum instead.

TWO QUESTIONABLE CLASSES. In addition to the seven formal classes just described, there are two other somewhat questionable classes of nonsilicate that might be included in a listing of minerals. They would be included, if at all, only with major reservations, since they do not strictly fit the fourfold definition of a mineral as crystalline in structure, natural, inorganic, and identifiable by a precise chemical formula. These two questionable groups are organics and mineraloids.

Organics, as their name suggests, have organic components, but as we have observed, "organic" is not the same as "biological." This class excludes hard substances created in a biological setting—for example, bone or pearl—and includes only minerals that develop in a geologic setting yet have organic chemicals in their composition. By far the best-known example of this class, which includes only a half-dozen minerals, is amber, which is fossilized tree sap.

Amber is also among the mineraloids, which are not really "questionable" at all—they are clearly *not* minerals, since they do not have the necessary crystalline structure. Nevertheless, they often are listed among minerals in reference books and are likely to be sold by mineral dealers. The other four mineraloids include two other well-known substances, opal and obsidian.

Silicates

Where minerals are concerned, the silicates are the "stars of the show": the most abundant and most widely used class of minerals. That being said, it should be pointed out that there are a handful of abundant nonsilicates, most notably the iron oxides hematite, magnetite, and goethite. A few other nonsilicates, while they are less abundant, are important to the makeup of Earth's crust, examples being the carbonates calcite and dolomite; the sulfides pyrite, sphalerite, galena, and chalcopyrite; and the sulfate gypsum. Yet the nonsilicates are not nearly as important as the class of minerals built around the element silicon.

Though it was discovered by Jöns Berzelius in 1823, owing to its abundance in the planet's minerals, silicon has been in use by humans for thousands of years. Indeed, silicon may have been one of the first elements formed in the Precambrian eons (see Geologic Time). Geologists believe that Earth once was composed primarily of molten iron, oxygen, silicon, and aluminum, which, of course, are still the predominant elements in the planet's crust. But because iron has a greater atomic mass, it settled toward the center, while the more lightweight elements rose to the surface. After oxygen, silicon is the most abundant of all elements on the planet, and compounds involving the two make up about 90% of the mass of Earth's crust.

SILICON, CARBON, AND OXYGEN. On the periodic table, silicon lies just below carbon, with which it shares an ability to form long strings of atoms. Because of this and other chemical characteristics, silicon, like carbon, is at the center of a vast array of compounds—organic in the case of carbon and inorganic in the case of silicon. Silicates, which, as noted earlier, account for nine-tenths of the mass of Earth's crust (and 30% of all minerals), are to silicon and mineralogy what hydrocarbons are to carbon and organic chemistry.

Whereas carbon forms it most important compounds with hydrogen—hydrocarbons such as petroleum—the most important silicon-containing compounds are those formed by bonds with oxygen. There is silica (SiO_2), for instance, commonly known as sand. Aside from its many applications on the beaches of the world, silica, when mixed with lime and soda (sodium carbonate) and other substances, makes glass. Like carbon, silicon has the ability to form polymers, or long, chainlike molecules. And whereas carbon polymers are built of hydrocarbons (plastics are an example), silicon polymers are made of silicon and oxygen in monomers, or strings of atoms, that form ribbons or sheets many millions of units long.

SILICATE SUBCLASSES. There are six subclasses of silicate, differentiated by structure. Nesosilicates include some the garnet group; gadolinite, which played a significant role in the isolation of the lanthanide series of elements during the nineteenth century; and zircon. The latter may seem to be associated with the cheap diamond simulant, or substitute, called cubic zirconium, or CZ. CZ, however, is an artificial "mineral," whereas zircon is the real thing—yet it, too, has been applied as a diamond simulant.

Just as silicon's close relative, carbon, can form sheets (this is the basic composition of graphite), so silicon can appear in sheets as the phyllosilicate subclass. Included among this group are minerals known for their softness: kaolinite, talc, and various types of mica. These are used in everything from countertops to talcum powder. The kaolinite derivative known as kaolin is applied, for instance, in the manufacture of porcelain, while some people in parts of Georgia, a state noted for its kaolinite deposits, claim that it can and should be chewed as an antacid stomach remedy. (One can even find little bags of kaolin sold for this purpose at convenience stores around Columbus in southern Georgia.)

CALCITE WITH QUARTZ. *(© Mark A. Schneider/Photo Researchers. Reproduced by permission.)*

Included in another subclass, the tectosilicates, are the feldspar and quartz groups, which are the two most abundant types of mineral in Earth's crust. Note that these are both groups: to a mineralogist, feldspar and quartz refer not to single minerals but to several within a larger grouping. Feldspar, whose name comes from the Swedish words for "field" and "mineral" (a reference to the fact that miners and farmers found the same rocks in their respective areas of labor), includes a number of varieties, such as albite (sodium aluminum silicate) or sanidine (potassium aluminum silicate).

Other, more obscure silicate subclasses include sorosilicates and inosilicates. Finally, there are cyclosilicates, such as beryl or beryllium aluminum silicate.

IDENTIFYING MINERALS

Mineralogists identify unknown minerals by judging them in terms of various physical properties, including hardness, color and streak, luster, cleavage and fracture, density and specific gravity, and other factors, such as crystal form. Hardness, or the ability of one mineral to scratch another, may be measured against the Mohs scale, introduced in 1812 by the German mineralogist Friedrich Mohs (1773–1839). The scale rates minerals from 1 to 10, with 10 being equivalent to the hardness of a diamond and 1 that of talc, the softest mineral. (See Economic Geology for other scales, some of which are more applicable to specific types of minerals.)

Minerals sometimes can be identified by color, but this property can be so affected by the presence of impurities that mineralogists rely instead on streak. The latter term refers to the color of the powder produced when one mineral is scratched by another, harder one. Another visual property is luster, or the appearance of a mineral when light reflects off its surface. Among the terms used in identifying luster are metallic, vitreous (glassy), and dull.

The term cleavage refers to the way in which a mineral breaks—that is, the planes across which the mineral splits into pieces. For instance, muscovite tends to cleave only in one direction, forming thin sheets, while halite cleaves in three directions, which are all perpendicular to one another, forming cubes. The cleavage of a mineral reveals its crystal system; however, minerals are more likely to fracture (break along something other than a flat surface) than they are to cleave.

DENSITY, SPECIFIC GRAVITY, AND OTHER PROPERTIES. Density is the ratio of mass to volume, and specific gravity is the ratio between the density of a particular substance and that of water. Specific gravity almost always is measured according to the metric system, because of the convenience: since the density of water is 1 g per cubic centimeter (g/cm^3), the specific gravity of a substance is identical to its density, except that specific gravity involves no units.

For example, gold has a density of 19.3 g/cm^3 and a specific gravity of 19.3. Its specific gravity, incidentally, is extremely high, and, indeed, one of the few metals that comes close is lead, which has a specific gravity of 11. By comparing specific gravity values and measuring the displacement of water when an object is set down in it, it is possible to determine whether an item purported to be gold actually is gold.

In addition to these more common parameters for identifying minerals, it may be possible to identify certain ones according to other specifics. There are minerals that exhibit fluorescent or phosphorescent characteristics, for instance. The first term refers to objects that glow when viewed under ultraviolet light, while the second term describes those that continue to glow after being exposed to visible light for a short period of time. Some minerals are magnetic, while others are radioactive.

NAMING MINERALS. Chemists long ago adopted a system for naming compounds so as to avoid the confusion of proliferating common names. The only compounds routinely referred to by their common names in the world of chemistry are water and ammonia; all others are known according to chemical nomenclature that is governed by specific rules. Thus, for instance, NaCl is never "salt," but "sodium chloride."

Geologists have not been able to develop such a consistent means of naming minerals. For one thing, as noted earlier, two minerals may be different from each other yet include the same elements. Furthermore, it is difficult (unlike the case of chemical compounds) to give minerals names that provide a great deal of information regarding their makeup. Instead, most minerals are simply named after people (usually scientists) or the locale in which they were found.

Abrasives

The physical properties of minerals, including many of the characteristics we have just discussed, have an enormous impact on their usefulness and commercial value. Some minerals, such as diamonds and corundum, are prized for their hardness, while others, ranging from marble to the "mineral" alabaster, are useful precisely because they are soft. Others, among them copper and gold, are not just soft but highly malleable, and this property makes them particularly useful in making products such as electrical wiring.

Diamonds, corundum, and other minerals valued for their hardness belong to a larger class of materials called abrasives. The latter includes sandpaper, which of course is made from one of the leading silicate derivatives, sand. Sandstone and quartz are abrasives, as are numerous variants of corundum, such as sapphire and garnets.

In 1891, American inventor Edward G. Acheson (1856–1931) created silicon carbide, later sold under the trade name Carborundum, by heating a mixture of clay and coke (almost pure carbon). For 50 years, Carborundum was the second-hardest substance known, diamonds being the hardest. Today other synthetic abrasives, made from aluminum oxide, boron carbide, and boron nitride, have supplanted Carborundum in importance.

Corundum, from the oxides class of mineral, can have numerous uses. Extremely hard, corundum, in the form of an unconsolidated rock commonly called emery, has been used as an abrasive since ancient times. Owing to its very high melting point—even higher than that of iron—corundum also is employed in making alumina, a fireproof product used in furnaces and fireplaces. Though pure corundum is colorless, when combined with trace amounts of certain elements, it can yield brilliant colors: hence, corundum with traces of chromium becomes a red ruby, while traces of iron, titanium, and other elements yield varieties of sapphire in yellow, green, and violet as well as the familiar blue.

This brings up an important point: many of the minerals named here are valued for much more than their abrasive qualities. Many of the 16 minerals used as gemstones, including corundum (source of both rubies and sapphires, as we have noted), garnet, quartz, and of course diamond, happen to be abrasives as well. (See Economic Geology for the full list of precious gems.)

DIAMONDS. Diamonds, in fact, are so greatly prized for their beauty and their application in jewelry that their role as "working" minerals—not just decorations—should be emphasized. The diamonds used in industry look quite different from the ones that appear in jewelry. Industrial diamonds are small, dark, and cloudy in appearance, and though they have the same chemical properties as gem-quality diamonds, they are cut with functionality (rather than beauty) in mind. A diamond is hard, but brittle: in other words, it can be broken, but it is very difficult to scratch or cut a diamond—except with another diamond.

On the other hand, the cutting of fine diamonds for jewelry is an art, exemplified in the alluring qualities of such famous gems as the jewels in the British Crown or the infamous Hope Diamond in Washington, D.C.'s Smithsonian Institution. Such diamonds—as well as the diamonds on an engagement ring—are cut to refract or bend light rays and to disperse the colors of visible light.

Soft and Ductile Minerals

At the other end of the Mohs scale are an array of minerals valued not for their hardness, but for opposite qualities. Calcite, for example, is often used in cleansers because, unlike an abrasive (also used for cleaning in some situations), it will not scratch a surface to which it is applied. Calcite takes another significant form, that of marble, which is used in sculpture, flooring, and ornamentation because of its softness and ease in carving—not to mention its great beauty.

Gypsum, used in plaster of paris and wallboard, is another soft mineral with applications in building. Though, obviously, soft minerals are not much value as structural materials, when stud walls of wood provide the structural stability for gypsum sheet wall coverings, the softness of the latter can be an advantage. Gypsum wallboard makes it easy to put in tacks or nails for pictures and other decorations, or to cut out a hole for a new door, yet it is plenty sturdy if bumped. Furthermore, it is much less expensive than most materials, such as wood paneling, that might be used to cover interior walls.

GOLD. Quite different sorts of minerals are valued not only for their softness but also their ductility or malleability. There is gold, for instance, the most ductile of all metals. A single troy ounce (31.1 g) can be hammered into a sheet just 0.00025 in. (0.00064 cm) thick, covering 68 sq. ft. (6.3 sq m), while a piece of gold weighing about as much as a raisin (0.0022 lb., or 1 g) can be pulled into the shape of a wire 1.5 mi. (2.4 km) long. This, along with its qualities as a conductor of heat and electricity, would give it a number of other applications, were it not for the high cost of gold.

Therefore, gold, if it were a person, would have to be content with being only the most prized and admired of all metallic minerals, an element for which men and whole armies have fought and sometimes died. Gold is one of the few metals that is not silver, gray, or white in appearance, and its beautifully distinctive color caught the eyes of metalsmiths and royalty from the beginning of civilization. Hence it was one of the first widely used metals.

Records from India dating back to 5000 B.C. suggest a familiarity with gold, and jewelry found in Egyptian tombs indicates the use of sophisticated techniques among the goldsmiths of Egypt as early as 2600 B.C. Likewise, the Bible mentions gold in several passages, and the Romans called it *aurum* ("shining dawn"), which explains its chemical symbol, Au.

COPPER. Copper, gold, and silver are together known as coinage metals. They have all been used for making coins, a reflection not only of their attractiveness and malleability, but also of their resistance to oxidation. (Oxygen has a highly corrosive influence on metals, causing rust, tarnishing, and other effects normally associated with aging but in fact resulting from the reaction of metal and oxygen.) Of the three coinage metals, copper is by far the most versatile, widely used for electrical wiring and in making cookware. Due to the high conductivity of copper, a heated copper pan has a uniform temperature, but copper pots must be coated with tin because too much copper in food is toxic.

Its resistance to corrosion makes copper ideal for plumbing. Likewise, its use in making coins resulted from its anticorrosive qualities, combined with its beauty. These qualities led to the use of copper in decorative applications for which gold would have been much too expensive: many old buildings used copper roofs, and the Statue of Liberty is covered in 300 thick copper plates. As for why the statue and many old copper roofs are green rather than copper-colored, the reason is that copper does eventually corrode when exposed to air for long periods of time. It develops a thin layer of black copper oxide, and as the years pass, the reaction with carbon dioxide in the air leads to the formation of copper carbonate, which imparts a greenish color.

Unlike silver and gold, copper is still used as a coinage metal, though it, too, has been increasingly taken off the market for this purpose due to the high expense involved. Ironically, though most people think of pennies as containing copper, in fact the penny is the only American coin that contains *no* copper alloys. Because the amount of copper necessary to make a penny today costs more than one cent, a penny is actually made of zinc with a thin copper coating.

Insulation and Other Applications

Whereas copper is useful because it conducts heat and electricity well, other minerals (e.g., kyanite, andalusite, muscovite, and silimanite) are valuable for their ability *not* to conduct heat or electricity. Muscovite is often used for insulation in electrical devices, though its many qualities make it a mineral prized for a number of reasons.

Its cleavage and lustrous appearance, combined with its transparency and almost complete lack of color, made it useful for glass in the windowpanes of homes owned by noblemen and other wealthy Europeans of the Middle Ages. Today, muscovite is the material in furnace and stove doors: like ordinary glass, it makes it possible for one to look inside without opening the door, but unlike glass, it is an excellent insulator. The glass-like quality of muscovite also makes it a popular material in wallpaper, where ground muscovite provides a glassy sheen.

In the same vein, asbestos—which may be made of chrysotile, crocidolite, or other minerals—has been prized for a number of qualities, including its flexibility and fiber-like cleavage. These factors, combined with its great heat resistance and its resistance to flame, have made it useful for fireproofing applications, as for instance in roofing materials, insulation for heating and electrical devices, brake linings, and suits for firefighters and others who must work around flames and great heat. However, information linking asbestos and certain forms of cancer, which began to circulate in the 1970s, led to a sharp decline in the asbestos industry.

MINERALS FOR HEALTH OR OTHERWISE. All sorts of other properties give minerals value. Halite, or table salt, is an important—perhaps too important!—part of the American diet. Nor is it the only consumable mineral; people also take minerals in dietary supplements, which is appropriate since the human body itself contains numerous minerals. In addition to a very high proportion of carbon, the body also contains a significant amount of iron, a critical component in red blood cells, as well as smaller amounts of minerals such as zinc. Additionally, there are trace minerals, so called because only traces of them are present in the body, that include cobalt, copper, manganese, molybdenum, nickel, selenium, silicon, and vanadium.

One mineral that does not belong in the human body is lead, which has been linked with a number of health risks. The human body can only excrete very small quantities of lead a day, and this is particularly true of children. Even in small concentrations, lead can cause elevation of blood pressure, and higher concentrations can effect the central nervous system, resulting in decreased mental functioning, hearing damage, coma, and possibly even death.

The ancient Romans, however, did not know this, and used what they called *plumbum* in making water pipes. (The Latin word is the root of our own term *plumber.*) Many historians believe that *plumbum* in the Romans' water supply was one of the reasons behind the decline and fall of the Roman Empire.

Even in the early twentieth century, people did not know about the hazards associated with lead, and therefore it was applied as an ingredient in paint. In addition, it was used in water pipes, and as an antiknock agent in gasolines. Increased awareness of the health hazards involved have led to a discontinuation of these practices.

GRAPHITE. Pencil "lead," on the other hand, is actually a mixture of clay with graphite, a form of carbon that is also useful as a dry lubricant because of its unusual cleavage. It is slippery because it is actually a series of atomic sheets, rather like a big, thick stack of carbon paper: if the stack is heavy, the sheets are likely to slide against one another.

Actually, people born after about 1980 may have little experience with carbon paper, which was gradually phased out as photocopiers became cheaper and more readily available. Today, carbon paper is most often encountered

KEY TERMS

ALLOY: A mixture of two or more metals.

ANION: The negative ion that results when an atom or group of atoms gains one or more electrons.

ATOM: The smallest particle of an element, consisting of protons, neutrons, and electrons. An atom can exist either alone or in combination with other atoms in a molecule.

ATOMIC NUMBER: The number of protons in the nucleus of an atom. Since this number is different for each element, elements are listed on the periodic table in order of atomic number.

CATION: The positive ion that results when an atom or group of atoms loses one or more electrons.

CHEMICAL BONDING: The joining, through electromagnetic forces, of atoms representing different elements. The principal methods of combining are through covalent and ionic bonding, though few bonds are purely one or the other.

CLEAVAGE: A term referring to the characteristic patterns by which a mineral breaks and specifically to the planes across which breaking occurs.

COMPOUND: A substance made up of atoms of more than one element, chemically bonded to one another.

COVALENT BONDING: A type of chemical bonding in which two atoms share valence electrons.

CRUST: The uppermost division of the solid earth, representing less than 1% of its volume and varying in depth from 3 mi. to 37 mi. (5–60 km).

CRYSTALLINE SOLID: A type of solid in which the constituent parts have a simple and definite geometric arrangement that is repeated in all directions.

ELECTRON: A negatively charged particle in an atom, which spins around the nucleus.

ELECTRONEGATIVITY: The relative ability of an atom to attract valence electrons.

ELEMENT: A substance made up of only one kind of atom. Unlike compounds, elements cannot be broken chemically into other substances.

HARDNESS: In mineralogy, the ability of one mineral to scratch another. This is measured by the Mohs scale.

HYDROCARBON: Any chemical compound whose molecules are made up of nothing but carbon and hydrogen atoms.

ION: An atom or group of atoms that has lost or gained one or more electrons and thus has a net electric charge. Positively charged ions are called *cations,* and negatively charged ones are called *anions.*

IONIC BONDING: A form of chemical bonding that results from attractions between ions with opposite electric charges.

when signing a credit-card receipt: the signature goes through the graphite-based backing of the receipt onto a customer copy.

In such a situation, one might notice that the copied image of the signature looks as though it were signed in pencil, which of course is fitting

KEY TERMS CONTINUED

LUSTER: The appearance of a mineral when light reflects off its surface. Among the terms used in identifying luster are metallic, vitreous (glassy), and dull.

MINERAL: A naturally occurring, typically inorganic substance with a specific chemical composition and a crystalline structure. Unknown minerals usually can be identified in terms of specific parameters, such as hardness or luster.

MINERALOGY: The study of minerals, which includes a number of smaller subdisciplines, such as crystallography.

MIXTURE: A substance with a variable composition, meaning that it is composed of molecules or atoms of differing types and in variable proportions.

MOHS SCALE: A scale, introduced in 1812 by the German mineralogist Friedrich Mohs (1773–1839), that rates the hardness of minerals from 1 to 10. Ten is equivalent to the hardness of a diamond and 1 that of talc, an extremely soft mineral.

MONOMERS: Small, individual subunits that join together to form polymers.

NUCLEUS: The center of an atom, a region where protons and neutrons are located and around which electrons spin.

ORE: A rock or mineral possessing economic value.

ORGANIC: At one time, chemists used the term *organic* only in reference to living things. Now the word is applied to most compounds containing carbon and hydrogen, thus excluding carbonates (which are minerals) and oxides such as carbon dioxide.

PERIODIC TABLE OF ELEMENTS: A chart that shows the elements arranged in order of atomic number, along with the chemical symbol and the average atomic mass for that particular element.

POLYMERS: Large, typically chainlike molecules composed of numerous smaller, repeating units known as monomers.

PROTON: A positively charged particle in an atom.

PURE SUBSTANCE: A substance, whether an element or compound, that has the same chemical composition throughout. Compare with *mixture.*

REACTIVITY: A term referring to the ability of one element to bond with others. The higher the reactivity (and, hence, the electronegativity value), the greater the tendency to bond.

ROCK: An aggregate of minerals.

SPECIFIC GRAVITY: The ratio between the density of a particular substance and that of water.

STREAK: The color of the powder produced when one mineral is scratched by another, harder one.

VALENCE ELECTRONS: Electrons that occupy the highest principal energy level in an atom. These are the electrons involved in chemical bonding.

due to the application of graphite in pencil "lead." In ancient times, people did indeed use lead—which is part of the "carbon family" of elements, along with carbon and silicon—for writing, because it left gray marks on a surface. Even today, people still use the word "lead" in reference to pencils, much as they still refer to a galvanized steel roof with a zinc coating as a "tin roof."

(For more about minerals, see Rocks. The economic applications of both minerals and rocks are discussed in Economic Geology. In addition, Paleontology contains a discussion of fossilization, a process in which minerals eventually replace organic material in long-dead organisms.)

WHERE TO LEARN MORE

Atlas of Rocks, Minerals, and Textures (Web site). <http://www.geosci.unc.edu/Petunia/IgMetAtlas/mainmenu.html>.

Hurlbut, Cornelius, W. Edwin Sharp, and Edward Salisbury Dana. *Dana's Minerals and How to Study Them.* New York: John Wiley and Sons, 1997.

The Mineral and Gemstone Kingdom: Minerals A–Z (Web site). <http://www.minerals.net/mineral/>.

Minerals and Metals: A World to Discover. Natural Resources Canada (Web site). <http://www.nrcan.gc.ca/mms/school/e_mine.htm>.

Minerals by Name (Web site). <http://mineral.galleries.com/minerals/by_name.htm>.

Pough, Frederick H. *A Field Guide to Rocks and Minerals*. Boston: Houghton Mifflin, 1996.

Roberts, Willard Lincoln, Thomas J. Campbell, and George Robert Rapp. *Encyclopedia of Minerals.* New York: Van Nostrand Reinhold, 1990.

Sorrell, Charles A., and George F. Sandström. *A Field Guide and Introduction to the Geology and Chemistry of Rocks and Minerals.* New York: St. Martin's, 2001.

Symes, R. F. *Rocks and Minerals.* Illus. Colin Keates and Andreas Einsiedel. New York: Dorling Kindersley, 2000.

"USGS Minerals Statistics and Information." United States Geological Survey (Web site). <http://minerals.usgs.gov/minerals/>.

ROCKS

CONCEPT

It might come as a surprise to learn that geologists regularly use an unscientific-sounding term, rocks. Yet as is almost always the case with a word used both in everyday language and within the realm of a scientific discipline, the meanings are not the same. For one thing, *rock* and *stone* are not interchangeable, as they are in ordinary discussion. The second of these two terms is used only occasionally, primarily as a suffix in the names of various rocks, such as limestone or sandstone. On the other hand, a rock is an aggregate of minerals or organic material. Rocks are of three different types: igneous, formed by crystallization of molten minerals, as in a volcano; sedimentary, usually formed by deposition, compaction, or cementation of weathered rock; and metamorphic, formed by alteration of preexisting rock.

HOW IT WORKS

An Introduction to Rocks

To expand somewhat on the definition of *rock,* the term may be said to describe an aggregate of minerals or organic material, which may or may not appear in consolidated form. Consolidation, which we will explore further within the context of sedimentary rock, is a process whereby materials become compacted, or experience an increase in density. It is likely that the image that comes to mind when the word rock is mentioned is that of a consolidated one, but it is important to remember that the term also can apply to loose particles.

The role of organic material in forming rocks also belongs primarily within the context of sedimentary, as opposed to igneous or metamorphic, rocks. There are, indeed, a handful of rocks that include organic material, an example being coal, but the vast majority are purely inorganic in origin. The inorganic materials that make up rocks are minerals, discussed in the next section. Rocks and minerals of economic value are called ores, which are examined in greater depth elsewhere, within the context of Economic Geology.

Minerals Defined

The definition of a mineral includes four components: it must appear in nature and therefore not be artificial, it must be inorganic in origin, it must have a definite chemical composition, and it must have a crystalline internal structure. The first of these stipulations clearly indicates that there is no such thing as a man-made mineral; as for the other three parts of the definition, they deserve a bit of clarification.

At one time, the term *organic,* even within the realm of chemistry, referred to all living or formerly living things, their parts, and substances that come from them. Today, however, chemists use the word to describe any compound that contains carbon and hydrogen, thus excluding carbonates (which are a type of mineral) and oxides such as carbon dioxide or carbon monoxide.

NONVARYING COMPOSITION. The third stipulation, that a mineral must be of nonvarying composition, limits minerals almost exclusively to elements and compounds—that is, either to substances that cannot be chemically broken down to yield simpler substances or to substances formed by the chemical bonding of elements. The chemical bonding of elements is a process quite different from mixing, and a compound is not to be confused with a mixture, whose composition is highly variable.

Another way of putting this is to say that all minerals must have a definite chemical formula, which is not possible with a mixture such as dirt or glass. The Minerals essay, which the reader is encouraged to consult for further information, makes reference to certain alloys, or mixtures of metals, that are classified as minerals. These alloys, however, are exceptional and fit certain specific characteristics of interest to mineralogists. The vast majority of the more than 3,700 known varieties of mineral constitute either a single element or a single compound.

CRYSTALLINE STRUCTURE. The fact that a mineral must have a crystalline structure implies that it must be a solid, since all crystalline substances are solids. A solid, of course, is a type of matter whose particles, in contrast to those of a gas or liquid, maintain an orderly and definite arrangement and resist attempts at compression. Thus, petroleum cannot be a mineral, nor is "mineral spirits," a liquid paint thinner made from petroleum (and further disqualified by the fact that it is artificial in origin).

Crystalline solids are those in which the constituent parts are arranged in a simple, definite geometric pattern that is repeated in all directions. These solids are contrasted with amorphous solids, such as clay. Metals are crystalline in structure; indeed, several metallic elements that appear on Earth in pure form (for example, gold, copper, and silver) also are classified as minerals.

Identifying Minerals

The type of crystal that appears in a mineral is one of several characteristics that make it possible for a mineralogist to identify an unidentified mineral. Although, as noted earlier, there are nearly 4,000 known varieties of mineral, there are just six crystal systems, or geometric shapes formed by crystals. Crystallographers, or mineralogists concerned with the study of crystal structures, are able to identify the crystal system by studying a good, well-formed specimen of a mineral, observing the faces of the crystal and the angles at which they meet.

Other characteristics by which minerals can be studied and identified visually are color, streak, and luster. The first of these features is not particularly reliable, because impurities in the mineral may greatly affect its hue. Therefore, mineralogists are much more likely to rely on streak, or the color of the powder produced when one mineral is scratched by a harder one. Luster, the appearance of a mineral when light reflects off its surface, is described by such terms as vitreous (glassy), dull, or metallic.

HARDNESS. Minerals also can be identified according to what might be called tactile properties, or characteristics best discerned through the sense of touch. One of the most important among such properties is hardness, defined as the ability of one mineral to scratch another. Hardness is measured by the Mohs scale, introduced in 1812 by the German mineralogist Friedrich Mohs (1773–1839).

The scale rates minerals from 1 to 10, with 1 being equivalent to the hardness of talc, a mineral so soft that it is used for making talcum powder. A 2 on the Mohs scale is the hardness of gypsum, which is still so soft that it can be scratched by a human fingernail. Above a 5 on the scale, roughly equal to the hardness of a pocketknife or glass, are potassium feldspar (6), quartz (7), topaz (8), corundum (9), and diamond (10).

OTHER PROPERTIES. Other tactile parameters are cleavage, the planes across which the mineral breaks, and fracture, the tendency to break along something other than a flat surface. Minerals also can be evaluated by their density (ratio of mass to volume) or specific gravity (ratio between the mineral's density and that of water). Density and specific-gravity measures are particularly important for extremely dense materials, such as lead or gold.

In addition to these specifics, others may be used for identifying some kinds of minerals. Magnetite and a few other minerals, for instance, are magnetic, while minerals containing uranium and other elements with a high atomic number may be radioactive, or subject to the spontaneous emission of high-energy particles. Still others are fluorescent, meaning that they glow when viewed under ultraviolet light, or phosphorescent, meaning that they continue to glow after being exposed to visible light for a short period of time.

Mineral Groups

Minerals are classified into eight basic groups:

- Class 1: Native elements
- Class 2: Sulfides
- Class 3: Oxides and hydroxides
- Class 4: Halides
- Class 5: Carbonates, nitrates, borates, iodates
- Class 6: Sulfates, chromates, molybdates, tungstates
- Class 7: Phosphates, arsenates, vanadates
- Class 8: Silicates

The first group, native elements, includes metallic elements that appear in pure form somewhere on Earth; certain metallic alloys, alluded to earlier; and native nonmetals, semimetals, and minerals with metallic and nonmetallic elements. Sulfides include the most important ores of copper, lead, and silver, while halides are typically soft and transparent minerals containing at least one element from the halogens family: fluorine, chlorine, iodine, and bromine. (The most well known halide, table salt, is a good example of an unconsolidated mineral.)

Oxides are noncomplex minerals that contain either oxygen or hydroxide (OH). Included in the oxide class are such well-known materials as magnetite and corundum, widely used in industry. Other nonsilicates (a term that stresses the importance of silicates among mineral classes) include carbonates, or carbon-based minerals, as well as phosphates and sulfates. The latter are distinguished from sulfides by virtue of the fact that they include a complex anion (a negatively charged atom or group of atoms) in which an atom of sulfur, chromium, tungsten, selenium, tellurium, or molybdenum (or a combination of these) is attached to four oxygen atoms.

There are two other somewhat questionable classes of nonsilicate that might be included in a listing of minerals—organics and mineraloids. Though they have organic components, organics—for example, amber—originated in a geologic and not a biological setting. Mineraloids, among them, opal and obsidian, are not minerals because they lack the necessary crystalline structure, but they can be listed under the more loosely defined heading of "rocks."

SILICATES. Only a few abundant or important minerals are nonsilicates, for example, the iron oxides hematite, magnetite, and goethite; the carbonates calcite and dolomite; the sulfides pyrite, sphalerite, galena, and chalcopyrite; and the sulfate gypsum. The vast majority of minerals, including the most abundant ones, belong to a single class, that of silicates, which accounts for 30% of all minerals. As their name implies, they are built around the element silicon, which bonds to four oxygen atoms to form what are called silica tetrahedra.

Silicon, which lies just below carbon on the periodic table of elements, is noted, like carbon, for its ability to form long strings of atoms. Carbon-hydrogen formations, or hydrocarbons, are the foundation of organic chemistry, while formations of oxygen and silicon—the two most abundant elements on Earth—provide the basis for a vast array of geologic materials. There is silica, for instance, better known as sand, which consists of silicon bonded to two carbon atoms.

Then there are the silicates, which are grouped according to structure into six subclasses. Among these subclasses, discussed in the Minerals essay, are smaller groupings that include a number of well-known mineral types: garnet, zircon, kaolinite, talc, mica, and the two most abundant minerals on Earth, feldspar and quartz. The name *feldspar* comes from the Swedish words *feld* ("field") and *spar* ("mineral"), because Swedish miners tended to come across the same rocks that Swedish farmers found themselves extracting from their fields.

REAL-LIFE APPLICATIONS

Rocks and Human Existence

Rocks are all around us, especially in our building materials but also in everything from jewelry to chalk. Then, of course, there are the rocks that exist in nature, whether in our backyards or in some more dramatic setting, such as a national park or along a rugged coastline. Indeed, humans have a long history of involvement with rocks—a history that goes far back to the aptly named Stone Age.

CHICHÉN-ITZÁ, A MAYAN STONE PYRAMID IN THE STATE OF YUCATÁN, MEXICO. *(© Ulrike Welsch/Photo Researchers. Reproduced by permission.)*

The latter term refers to a period in which the most sophisticated human tools were those made of rock—that is, before the development of the first important alloy used in making tools, bronze. The Bronze Age began in the Near East in about 3300 B.C. and lasted until about 1200 B.C., when the development of iron-making technology introduced still more advanced varieties of tools.

These dates apply to the Near East, specifically to such areas as Mesopotamia and Egypt, which took the lead in ancient technology, followed much later by China and the Indus Valley civilization of what is now Pakistan. The rest of the world was even slower in adopting the use of metal: for instance, the civilizations of the Americas did not enter the Bronze Age for almost 4,000 years, in about A.D. 1100. Nor did they ever develop iron tools before the arrival of the Europeans in about 1500.

THE STONE AGE. In any case, the Stone Age, which practically began with the species *Homo sapiens* itself, was unquestionably the longest of the three ages. The Stone Age is divided into two periods: Paleolithic and Neolithic, sometimes called Old and New Stone Age, respectively. (There was also a middle phase, called the Mesolithic, but this term is not used as widely as Paleolithic or Neolithic.) Throughout much of this time, humans lived in rock caves and used rock tools, including arrowheads for killing animals and (relatively late in prehistory) flint for creating fire.

The Paleolithic, characterized by the use of crude tools chipped from pieces of stone, began sometime between 2.5 and 1.8 million years ago and lasted until last ice age ended (and the present Holocene epoch began), about 10,000 years ago. The Neolithic period that followed saw enormous advances in technology, so many advances that historians speak of a "Neolithic Revolution" that included the development of much more sophisticated, polished tools. The mining of gold, copper, and various other ores began long before the development of the first alloys (bronze is formed by the mixture of copper and tin). Yet even after humans discovered metals, they continued to use stone tools.

THE PYRAMIDS AND OTHER STONE STRUCTURES. Indeed, the great pyramids of Egypt, built during the period from about 2600-2400 B.C., were constructed primarily with the use of stone rather than metal tools. The structures themselves, of course, also reflect the tight connection between humans and rocks. Built of limestone, the pyramids are still standing some 4,500 years later, even as structures of clay and mud built at about the same time in Mesopotamia (a region poor in stone resources) have long since dwindled to dust.

Incidentally, the great pyramids once had surfaces of polished limestone, such that they gleamed in the desert sun. Centuries later, Arab invaders in the seventh century A.D. stripped this limestone facing to use it in other structures, and the only part of the facing that remains today is high atop the pyramid of Khafre. For this reason, Khafre's pyramid is slightly taller than the structure known as *the* Great Pyramid, that of Cheops, or Khufu, which was originally the largest pyramid.

The centuries that have followed the building of those great structures likewise are defined, at least in part, by their buildings of stone. The Bible is full of references to stones, whether those used in building Solomon's temple or the precious gemstones said to form the gates of the New Jerusalem described in the Book of Revelation. Greece and Rome, too, are known for their structures of stone, ranging from marble (limestone that has undergone metamorphism) to unconsolidated stones in early forms of concrete, pioneered by the Egyptians.

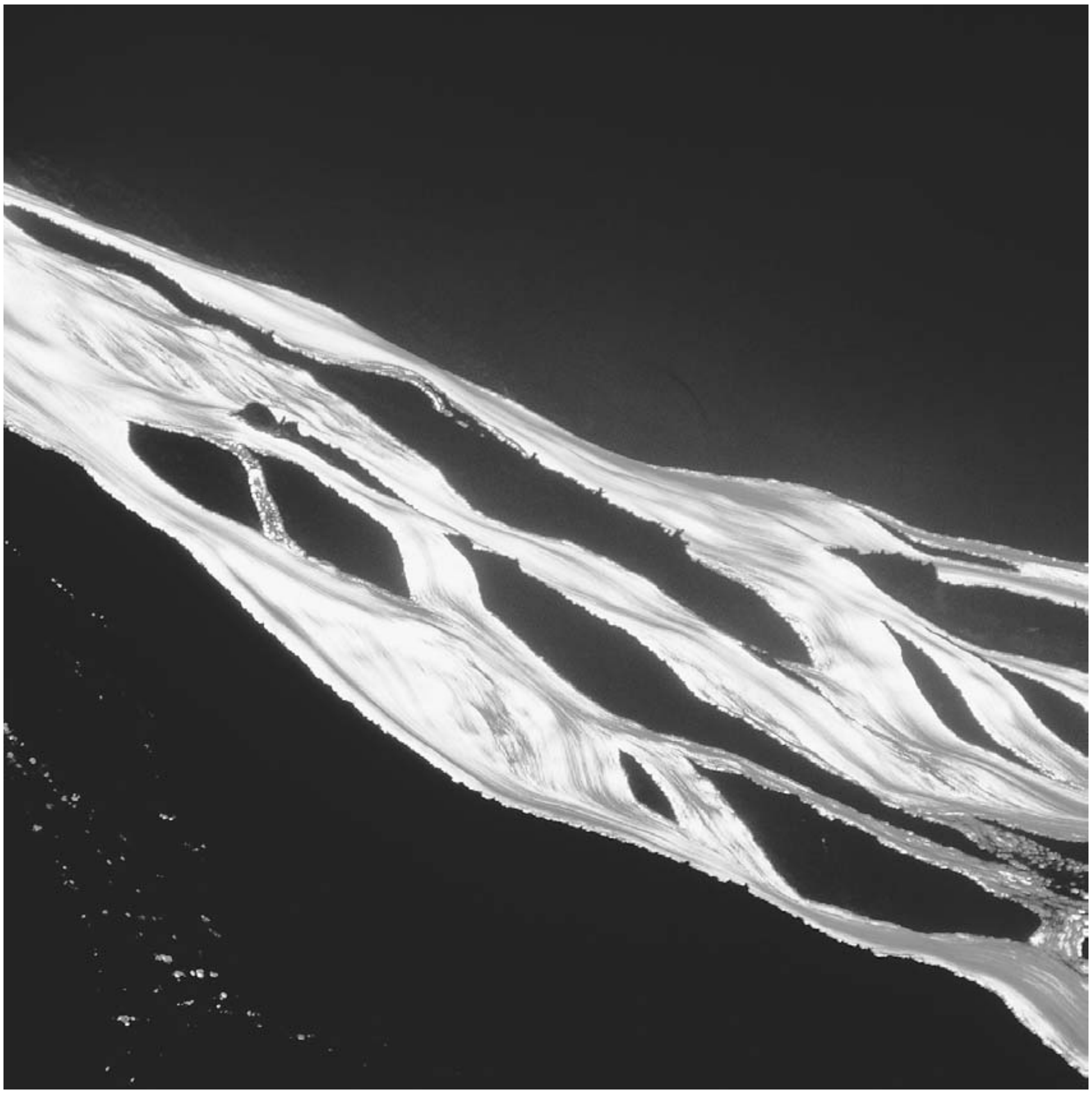
Still later, medieval Europe built its cathedrals and castles of stone, though it should be noted that the idea of the castle came from the Middle East, where the absence of lumber for fortresses caused Syrian castle builders to make use of abundant sandstone instead. Other societies left behind their own great stone monuments: the Great Wall of China, Angkor Wat in southeast Asia, the pyramids of Central America and Machu Picchu in South America, the great cliffside dwellings of what is now the southwestern United States, and the stone churches of medieval Ethiopia.

Certainly there were civilizations that created great structures of wood, but these structures were simply not as durable. The oldest wood building, a Buddhist temple at Horyuji in Japan, dates back only to A.D. 607, which, of course, is quite impressive for a wooden structure. But it hardly compares to what may well be the oldest known human structure, a windbreak discovered by the paleobiologist Mary Leakey (1913–1996) in Tanzania in 1960. Consisting of a group of lava blocks that form a rough circle, it is believed to be 1.75 million years old.

Mineralogy and Petrology

Not surprisingly, mineralogy is concerned with minerals—their physical properties, chemical makeup, crystalline structures, occurrence, distribution, and physical origins. Researchers whose work focuses on the physical origins of minerals study data and draw on the principles of physics and chemistry to develop hypotheses regarding the ways minerals form. Other mineralogical studies may involve the identification of a newly discovered mineral or the synthesis of mineral-like materials for industrial purposes.

The study of rocks is called petrology, from a Greek root meaning "rock." (Hence also the words *petroleum* and *petrify.*) Its areas of interest with regard to rocks are much the same as those of mineralogy as they relate to minerals: physical properties, distribution, and origins. It includes two major subdisciplines, experimental petrology, or the synthesis of rocks in a laboratory as a means of learning the conditions under which rocks are formed in the natural world, and petrography, or the study of rocks observed in thin sections through a petrographic microscope, which uses polarized light.



Lava flow after the 1992 eruption of Mount Etna in Sicily. When it cools, lava becomes igneous rock. *(© B. Edmaier/Photo Researchers. Reproduced by permission.)*

Owing to the fact that most rocks contain minerals, petrology draws on and overlaps with mineralogical studies to a great extent. At the same time, it goes beyond mineralogy, inasmuch as it is concerned with materials that contain organic substances, which are most likely to appear within the realm of sedimentary rock. Petrologists also are concerned with the other two principal types of rock, igneous and metamorphic.

Igneous Rocks

Igneous rock is rock formed by the crystallization of molten materials. It most commonly is associated with volcanoes, though, in fact, it comes into play in the context of numerous plate tectonic processes, such as seafloor spreading (see Plate Tectonics). The molten rock that becomes igneous rock is known as magma when it is below the surface of the earth and lava when it is at or near the earth's surface. Its most notable characteristic is its interlocking crystals. For the most part, igneous rocks do not have a layered texture.

When igneous rocks form deep within the Earth, they are likely to have large crystals, an indication of the fact that a longer period of time elapsed while the magma was cooling. On the other hand, volcanic rocks and others that form at or near Earth's surface are apt to have very small crystals. Obsidian (which, as we have noted, is not truly a mineral owing to its lack of crystals) is formed when hot lava comes into contact with water; as a result, it cools so quickly that crystals never have time to develop. Sometimes called volcanic glass, it once was used by prehistoric peoples as a cutting tool.

CLASSIFYING AND IDENTIFYING IGNEOUS ROCKS. Igneous rocks can be classified in several ways, referring to the means by which they were formed, the size of their crystals, and their mineral content. Extrusive igneous rocks, ejected by volcanoes to crystallize at or near Earth's surface, have small crystals, whereas intrusive igneous rocks, which cooled slowly beneath the surface, have larger crystals. Sometimes the terms *plutonic* and *volcanic,* which roughly correspond to *intrusive* and *extrusive,* respectively, are used.

Igneous rocks made of fragments from volcanic explosions are known as *pyroclastic,* or "fire-broken," rocks. Those that consist of dense, dark materials are known as *mafic* igneous rocks. On the other hand, those made of lightly colored, less-dense minerals, such as quartz, mica, and feldspar, are called *felsic* igneous rocks. Among the most well known varieties of igneous rock is granite, an intrusive, felsic rock that includes quartz, feldspar, mica, and amphibole in its makeup. Also notable is basalt, which is mafic and extrusive.

Sedimentary Rocks

Earlier, we touched on the subject of consolidation, which can be explained in more depth within the context of sedimentary rock. Consolidation is the compacting of loose materials by any number of processes, including recrystallization and cementation. The first of these processes is the formation of new mineral grains as a result of changes in temperature, pressure, or other factors. In cementation, particles of sediment (material deposited at or near Earth's surface from several sources, most notably preexisting rock) are cemented together, usually with mud.

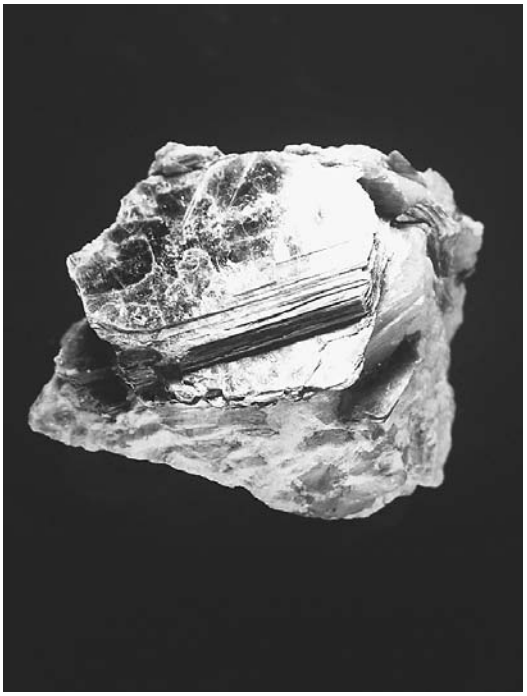
Compaction, recrystallization, and other processes, such as dehydration (which also may contribute to compaction), are collectively known as diagenesis. The latter term refers to all the changes experienced by a sediment sample under conditions of low temperature and low pressure following deposition. If the temperature and pressure increase, diagenesis may turn into metamorphism, discussed later in the context of metamorphic rock.

FORMATION OF SEDIMENTARY ROCKS. Sedimentary rock is formed by the deposition, compaction, and cementation of rock that has experienced weathering (breakdown of rock due to physical, chemical, or biological processes) or as a result of chemical precipitation. The latter term refers not to "precipitation" in terms of weather but to the formation of a solid from a liquid, by chemical rather than physical means. (The freezing of water, a physical process, is not an example of precipitation.)

Sedimentary rock usually forms at or near the surface of the earth, as the erosive action of wind, water, ice, gravity, or a combination of these forces moves sediment. Yet this formation also may occur when chemicals precipitate from seawater or when organic material, such as plant debris or animal shells, accumulate. Evaporation of saltwater, for instance, produces gypsum, a mineral noted for its lack of thermal conductivity; hence its use in drywall, the material that covers walls in most modern homes. (Ancient peoples made alabaster, a fine-grained ornamental stone, from gypsum.)

CLASSIFICATION AND SIZES. Sedimentary rock is classified with reference to the size of the particles from which the rock is made as well as the origin of those particles. Clastic rock comes from fragments of preexisting rock (whether igneous, sedimentary, or metamorphic) and organic matter, while nonclastic sedimentary rock is formed either by precipitation or by organic means. Examples include gypsum, salts, and other rocks formed by precipitation of saltwater as well as those created from organic material or organic activity—coal, for example.

Ranging in size from fine clay (less than 0.00015 in., or 0.004 mm) to boulders (defined as any rock larger than 10 in., or 0.254 m), sedimentary rock bears a record of the environment in which the original sediments were deposited. This record lies in the sediment itself. For example, rocks containing conglomerate, material ranging in size from clay to boulders (including the intermediate categories of silt, sand, gravel, pebbles, and cobble), come from sediment that was deposited rapidly as the result of slides or slumps. (Slides and slumps are discussed in Mass Wasting.)



METAMORPHIC ROCK IS FORMED BY THE ALTERATION OF PREEXISTING ROCK. THE PRESENCE OF MICA, SHOWN HERE, IS A SIGN THAT ROCK MIGHT BE METAPHORPHIC. *(© C. D. Winters/Photo Researchers. Reproduced by permission.)*

Sedimentary rocks are of particular interest to paleontologists, stratigraphers, and others working in the field of historical geology, because they are the only kinds of rock in which fossils are preserved. The pressure and temperature levels that produce igneous and metamorphic rock would destroy the organic remnants that produce fossils; on the other hand, sedimentary rock—created by much less destructive processes—permits the formation of fossils. Thus, the study of these formations has contributed greatly to geologists' understanding of the distant past. (See the essays Historical Geology, Stratigraphy, and Paleontology. For more about sedimentary rock, see Sediment and Sedimentation.)

METAMORPHIC ROCKS

Metamorphic rock is formed through the alteration of preexisting rock as a result of changes in temperature, pressure, or the activity of fluids (usually gas or water). These changes in temperature must be extreme (figures are given later), such that the preexisting rock—whether igneous, sedimentary, or metamorphic—is no longer stable.

Often formed in mountain environments, metamorphic rocks include such well-known varieties as marble, slate, and gneiss—metamorphosed forms of limestone, shale, and granite, respectively. Also notable is schist, composed of various minerals, such as talc, mica, and muscovite. There is not always a one-to-one correspondence between precursor rocks and metamorphic ones: increasing temperature and pressure can turn shale progressively into slate, phyllite, schist, and gneiss.

The presence of mica in a rock—or of other minerals, including amphibole, staurolite, and garnet—is a sign that the rock might be metamorphic. These minerals, typical of metamorphic rocks, are known as metamorphic facies. Also indicative of metamorphism are layers in the rock, more or less parallel lines along which minerals are laid as a result of the high pressures applied to the rock in its formation. Metamorphism, the process whereby metamorphic rock is created, also may produce characteristic formations, such as an alignment of elongate crystals or the separation of minerals into layers.

METAMORPHISM. Given the conditions described for metamorphism, one might conclude that in terms of violence, drama, and stress, it is a process somewhere between sedimentation and the formation of igneous rock. That, in fact, is precisely the case: the temperature and pressure conditions necessary for metamorphism lie between those of diagenesis, on the one hand, and the extreme conditions necessary for the production of igneous rock, on the other hand. Specifically, metamorphism occurs at temperatures between 392°F (200°C) and 1,472°F (800°C) and under levels of pressure between 1,000 and 10,000 bars. (A bar is slightly less than the standard atmospheric pressure at sea level. The latter, equal to 14.7 lb. per square inch, or 101,325 Pa, is equal to 1.01325 bars.)

There are several types of metamorphism: regional, contact, dynamic, and hydrothermal. Regional metamorphism results from a major tectonic event or events, producing widespread changes in rocks. Contact or thermal metamorphism results from contact between igneous intrusions and cooler rocks above them, which recrystallize as a result of heating. Dynamic metamorphism takes place in the high-pressure conditions along faults. Finally, hydrothermal metamorphism ensues from contact with fluids

KEY TERMS

ALLOY: A mixture of two or more metals.

CEMENTATION: A process of consolidation whereby particles of sediment are cemented together, usually with mud.

COMPOUND: A substance made up of atoms of more than one element, chemically bonded to one another.

CONGLOMERATE: Unconsolidated rock material containing rocks ranging in size from very small clay (less than 0.00015 in., or 0.004 mm) to boulders (defined as any rock larger than 10 in., or 0.254 m). Sedimentary rock often appears in the form of conglomerate.

CONSOLIDATION: A process whereby materials become compacted, or experience an increase in density. This takes place through several processes, including recrystallization and cementation.

CRYSTALLINE SOLID: A type of solid in which the constituent parts have a simple and definite geometric arrangement that is repeated in all directions.

DEPOSITION: The process whereby sediment is laid down on the Earth's surface.

DIAGENESIS: A term referring to all the changes experienced by a sediment sample under conditions of low temperature and low pressure following deposition. Higher temperature and pressure conditions may lead to metamorphism.

ELEMENT: A substance made up of only one kind of atom. Unlike compounds, elements cannot be broken chemically into other substances.

EROSION: The movement of soil and rock as the result of forces produced by water, wind, glaciers, gravity, and other influences. In most cases, a fluid medium, such as air or water, is involved.

IGNEOUS ROCK: One of the three principal types of rock, along with sedimentary and metamorphic rock. Igneous rock is formed by the crystallization of molten materials, for instance, in a volcano or other setting where plate tectonic processes take place.

LAVA: Molten rock at or near the surface of the earth that becomes igneous rock. Below the surface, lava is known as *magma.*

MAGMA: Molten rock beneath the surface of the earth that becomes igneous rock. Once it is at or near the surface, magma is known as *lava.*

MINERAL: A naturally occurring, typically inorganic substance with a specific chemical composition and a crystalline structure.

MINERALOGY: An area of geology devoted to the study of minerals. Mineralogy includes several subdisciplines, such as crystallography, the study of crystal formations within minerals.

MIXTURE: A substance with a variable composition, meaning that it is composed of molecules or atoms of differing types in varying proportions.

ORE: A rock or mineral possessing economic value.

ORGANIC: At one time, chemists used the term *organic* only in reference to living things. Now the word is applied to most

KEY TERMS CONTINUED

compounds containing carbon and hydrogen, thus excluding carbonates (which are minerals), and oxides such as carbon dioxide.

PETROLOGY: An area of geology devoted to the study of rocks, including their physical properties, distribution, and origins.

PRECIPITATION: In the context of chemistry, precipitation refers to the formation of a solid from a liquid.

RECRYSTALLIZATION: The formation of new mineral grains as a result of changes in temperature, pressure, or other factors.

ROCK: An aggregate of minerals or organic matter, which may be consolidated or unconsolidated.

ROCK CYCLE: The ongoing process whereby rocks continually change from one type to another, typically through melting, metamorphism, uplift, weathering, burial, or other processes.

SAND: A term that can have several meanings. The sand at a beach could be a variety of unconsolidated materials, though most likely it is silica (SiO_2). Sand is also a term used for a size of rock ranging from very fine to very coarse.

SEDIMENT: Material deposited at or near Earth's surface from a number of sources, most notably preexisting rock.

SEDIMENTARY ROCK: One of the three major types of rock, along with igneous and metamorphic rock. Sedimentary rock usually is formed by the deposition, compaction, and cementation of rock that has experienced weathering. It also may be formed as a result of chemical precipitation.

UNCONSOLIDATED ROCK: Rock that appears in the form of loose particles, such as sand.

UPLIFT: A process whereby the surface of Earth rises, owing to either a decrease in downward force or an increase in upward force.

WEATHERING: The breakdown of rocks and minerals at or near the surface of Earth as the result of physical, chemical, or biological processes.

heated by igneous rock. Reacting with minerals in the surrounding rock, the fluids produce different minerals, which, in turn, yield metamorphic rocks.

TYPES OF METAMORPHIC ROCKS. Metamorphic rocks that contain elongate or platy minerals, such as mica and amphibole, are called foliated *rocks.* These rocks have a layered texture, which may manifest as the almost perfect arrangement of materials in slate or as the alternating patterns of light and dark found in some other varieties of rock. Metamorphic rocks without visible layers are referred to as unfoliated rocks. As a foliated metamorphic rock, slate is particularly good for splitting into thin layers—hence one of its most important applications is in making shingles for roofing. By contrast, marble, which is unfoliated, is valued precisely for its lack of tendency to split.

Petrologists attempting to determine exactly which rocks or combinations of rocks metamorphosed to produce a particular sample often face a challenge. Many metamorphic rocks are stubborn about giving up their secrets; on the other hand, it is possible to match up precursor rocks with certain varieties. For example, as noted earlier, marble comes from limestone, while gneiss usually (but not always) comes from granite. Quartzite is metamorphosed sandstone. Nonetheless, it is not as easy to trace the history of a metamorphic rock as it is to say that a raisin was once a grape or that a pickle was once a cucumber.

Where to Find Rocks

In general, one might find igneous rocks such as basalt in any place known for volcanic activity either in the recent or distant past. This would include such well-known areas of volcanism as Hawaii, the Philippines, and Italy, but also places where volcanic activity occurred in the distant past. (See, for instance, the discussion in the essay titled "Paleontology" regarding possible volcanic activity in what is now the continental United States at the conclusion of the Triassic period.)

The best place for metamorphic rock would be in areas of mountain-building and powerful tectonic activity, as for instance in the Himalayas or the Alps of central Europe. Sedimentary rock is basically everywhere, but a good place to find large samples of it would include areas with large oil deposits, which are always found in sedimentary rock.

Closer to home, a wide array of sedimentary rocks can be located in the plains and lowlands of the United States, particularly in the West and Midwest, where large samples are exposed. Igneous and metamorphic rocks can be found, predictably, in regions where mountains provide evidence of past tectonic activity: New England, the Appalachians, and the various mountain ranges of the western United States such as the Rockies, Cascades, and Sierra Nevada.

The Rock Cycle

Given what we have seen about the characteristics of the three rock varieties—igneous, sedimentary, and metamorphic—it should be clear that there is no such thing as a rock that simply is what it is, without any possibility of changing. Rocks, in fact, are constantly changing, as is Earth itself. This process whereby rocks continually change from one type to another—typically through melting, metamorphism, uplift, weathering, burial, or other processes—is known as the rock cycle.

The rock cycle can go something like this: Exposed to surface conditions such as wind and the activity of water, rocks experience weathering. The result is the formation of sediments that are eventually compacted to make sedimentary rocks. As the latter are buried deeper and deeper beneath greater amounts of sediment, the pressure and temperature builds. This process ultimately can result in the creation of metamorphic rock. On the other hand, the rock may undergo such extreme conditions of temperature that it recrystallizes to form igneous rock. Whatever the variety—igneous, sedimentary, or metamorphic—the rock likely will be in a position eventually to experience erosion, in which case the rock cycle begins all over again.

Where to Learn More

Atlas of Igneous and Metamorphic Rocks, Minerals, and Textures (Web site). <http://www.geosci.unc.edu/Petunia/IgMetAtlas/mainmenu.html>.

Bishop, A. C., Alan Robert Woolley, and William Roger Hamilton. *Cambridge Guide to Minerals, Rocks, and Fossils.* New York: Cambridge University Press, 1999.

Busbey, Arthur Bresnahan. *Rocks and Fossils.* Alexandria, VA: Time-Life Books, 1996.

Discovery Channel Rocks and Minerals: An Explore Your World Handbook. New York: Discovery Books, 1999.

"The Essential Guide to Rocks." BBC Education (Web site). <http://www.bbc.co.uk/education/rocks/>.

"Igneous, Sedimentary, and Metamorphic Rock Info." University of British Columbia (Web site). <http://www.science.ubc.ca/~geol202/petrology/rock.html>.

RocksForKids.com (Web site). <http://www.rocksforkids.com/>.

Rocks and Minerals (Web site). <http://www.fi.edu/tfi/units/rocks/rocks.html>.

Symes, R. F., Colin Keates, and Andreas Einsiedel. *Rocks and Minerals.* New York: Dorling Kindersley, 2000.

Vernon, R. H. *Beneath Our Feet: The Rocks of Planet Earth.* New York: Cambridge University Press, 2000.

ECONOMIC GEOLOGY

CONCEPT

Economic geology is the study of fuels, metals, and other materials from the earth that are of interest to industry or the economy in general. It is concerned with the distribution of resources, the costs and benefits of their recovery, and the value and availability of existing materials. These materials include ore (rocks or minerals possessing economic value) as well as fossil fuels, which embrace a range of products from petroleum to coal. Rooted in several subdisciplines of the geologic sciences—particularly geophysics, structural geology, and stratigraphy—economic geology affects daily life in myriad ways. Masonry stones and gasoline, gypsum wallboard (sometimes known by the brand name Sheetrock) and jewelry, natural gas, and table salt—these and many more products are the result of efforts in the broad field known as economic geology.

HOW IT WORKS

Background of Economic Geology

Some sources of information in the geologic sciences use a definition of "economic geology" narrower than the one applied here. Rather than including nonmineral resources that develop in and are recovered from a geologic environment—a category that consists primarily of fossil fuels—this more limited definition restricts the scope of economic geology to minerals and ores. Given the obvious economic importance of fossil fuels such as petroleum and its many byproducts as well as coal and peat, however, it seems only appropriate to discuss these valuable organic resources alongside valuable inorganic ones.

The concept of economic geology as such is a relatively new one, even though humans have been extracting metals and minerals of value from the ground since prehistoric times. For all their ability to appreciate the worth of such resources, however, premodern peoples possessed little in the way of scientific theories regarding either their formation or the means of extracting them.

The Greeks, for instance, believed that veins of metallic materials in the earth indicated that those materials were living things putting down roots after the manner of trees. Astrologers of medieval times maintained that each of the "seven planets" (Sun, Moon, and the five planets, besides Earth, known at the time) ruled one of the seven known metals—gold, copper, silver, lead, tin, iron, and mercury—which supposedly had been created under the influence of their respective "planets."

AGRICOLA'S CONTRIBUTION. The first thinker who attempted to go beyond such unscientific (if imaginative) ideas was a German physician writing under the Latinized name Georgius Agricola (1494–1555). As a result of treating miners for various conditions, Agricola, whose real name was Georg Bauer, became fascinated with minerals. The result was a series of written works, culminating with *De re metallica* (On the nature of minerals, 1556, released postthumusly), that collectively initiated the modern subdiscipline of physical geology. (It is worth noting that the first translators of *Metallica* into English were Lou [d. 1944] and Herbert Clark Hoover [1874–1964]. The couple published their translation in 1912 in London's *Mining Magazine,* and the husband went on to become the thirty-first president of the United States in 1929.)

Rejecting the works of the ancients and all manner of fanciful explanations for geologic phenomena, Agricola instead favored careful observation, on the basis of which he formed verifiable hypotheses. Regarded as the father of both mineralogy and economic geology, Agricola introduced several ideas that provided a scientific foundation for the study of Earth and its products. In *De ortu et causis subterraneorum* (1546), he critiqued all preceding ideas regarding the formation of ores, including the Greek and astrological notions mentioned earlier as well as the alchemical belief that all metals are composed of mercury and sulfur. Instead, he maintained that subterranean fluids carry dissolved minerals, which, when cooled, leave deposits in the cracks of rocks and thus give rise to mineral veins. Agricola's ideas later helped form the basis for modern theories regarding the formation of ore deposits.

In *De natura fossilium* (On the nature of fossils, 1546), Agricola also introduced a method for the classification of "fossils," as minerals were then known. Agricola's system, which categorizes minerals according to such properties as color, texture, weight, and transparency, is the basis for the system of mineral classification in use today. Of all his works, however, the most important was *De re metallica,* which would remain the leading textbook for miners and mineralogists during the two centuries that followed. In this monumental work, he introduced many new ideas, including the concept that rocks contain ores that are older than the rocks themselves. He also explored in detail the mining practices in use during his time, itself an extraordinary feat in that miners of the sixteenth century tended to guard their trade secrets closely.

Metals, Minerals, and Rocks

Of all known chemical elements, 87, or about 80%, are metals. The latter group is identified as being lustrous or shiny in appearance and malleable or ductile, meaning that they can be molded into different shapes without breaking. Despite their ductility, metals are extremely durable, have high melting and boiling points, and are excellent conductors of heat and electricity. Some, though far from all, register high on the Mohs hardness scale, discussed later in the context of minerals.

The bonds that metals form with each other, or with nonmetals, are known as ionic bonds, the strongest type of chemical bond. Even within a metal, however, there are extremely strong, nondirectional bonds. Therefore, though it is easy to shape metals, it is very difficult to separate metal atoms. Obviously, most metal are solids at room temperature, though this is not true of all: mercury is liquid at ordinary temperatures, and gallium melts at just 85.6°F (29.76°C). Generally, however, metals would be described as crystalline solids, meaning that their constituent parts have a simple and definite geometric arrangement that is repeated in all directions. Crystalline structure is important also within the context of minerals as well as the rocks that contain them.

MINERALS. Whereas there are only 87 varieties of metal, there are some 3,700 types of mineral. There is considerable overlap between metals and minerals, but that overlap is far from complete: many minerals include nonmetallic elements, such as oxygen and silicon. A mineral is a substance that appears in nature and therefore cannot be created artificially, is inorganic in origin, has a definite chemical composition, and possesses a crystalline internal structure.

The term *organic* does not refer simply to substances with a biological origin; rather, it describes any compound that contains carbon, with the exception of carbonates (which are a type of mineral) and oxides, such as carbon dioxide or carbon monoxide. The fact that a mineral must be of nonvarying composition limits minerals almost exclusively to elements and compounds—that is, either to substances that cannot be broken down chemically to yield simpler substances or to substances formed by the chemical bonding of elements. Only in a few highly specific circumstances are naturally occurring alloys, or mixtures of metals, considered minerals.

MINERAL GROUPS. Minerals are classified into eight basic groups:

- Class 1: Native elements
- Class 2: Sulfides
- Class 3: Oxides and hydroxides
- Class 4: Halides

- Class 5: Carbonates, nitrates, borates, iodates
- Class 6: Sulfates, chromates, molybdates, tungstates
- Class 7: Phosphates, arsenates, vanadates
- Class 8: Silicates

The first group, native elements, includes metallic elements that appear in pure form somewhere on Earth; certain metallic alloys, alluded to earlier; as well as native nonmetals, semimetals, and minerals with metallic and nonmetallic elements. The native elements, along with the six classes that follow them in this list, are collectively known as nonsilicates, a term that emphasizes the importance of the eighth group. (For more about the nonsilicates, as well as other subjects covered in the present context, see Minerals.)

The vast majority of minerals, including the most abundant ones, belong to the silicates class, which is built around the element silicon. Just as carbon can form long strings of atoms, particularly in combination with hydrogen (as we discuss in the context of fossil fuels later in this essay), silicon also forms long strings, though its "partner of choice" is typically oxygen rather than hydrogen. Together with oxygen, silicon—known as a metalloid because it exhibits characteristics of both metals and nonmetals—forms the basis for an astonishing array of products, both natural and man-made, which we examine in brief later.

CHARACTERISTICS OF MINERALS. From the list of parameters first developed by Agricola has grown a whole array of characteristics by which minerals are classified. These characteristics also can be used to evaluate an unknown mineral and thus to determine the mineral class within which it fits. One such parameter is the type of crystal of which a mineral is composed. Though there are thousands of minerals, there are just six crystal systems, or basic geometric shapes formed by crystals. Crystallographers, mineralogists concerned with the study of crystal structures, are able to identify the crystal system (the simplest being isometric, or cubic) by studying a good specimen of a mineral and observing the faces of the crystal and the angles at which they meet.

Minerals also can be identified by their hardness, defined as the ability of one mineral to scratch another. Hardness can be measured by the Mohs scale, introduced in 1812 by the German mineralogist Friedrich Mohs (1773–1839), which rates minerals from 1 (talc) to 10 (diamond.) Though it is useful for geologists attempting to identify a mineral in the field, the Mohs scale is not considered helpful for the industrial testing of fine-grained materials, such as steel or ceramics. For such purposes, the Vickers or Knoop scales are applied. These scales (named, respectively, after a British company and an American official) also have an advantage over Mohs in that they offer a precise, proportional scale in which each increase of number indicates the same increase in hardness. By contrast, on the Mohs scale, an increase from 3 to 4 (calcite to fluorite) indicates an additional 25% in hardness, whereas a shift from 9 to 10 (corundum to diamond) marks an increase of 300%.

Other properties significant in identifying minerals are color; streak, or the appearance of the powder produced when one mineral is scratched by a harder one; luster, the appearance of a mineral when light reflects off its surface; cleavage, the planes across which a mineral breaks; fracture, the tendency to break along something other than a flat surface; density, or ratio of mass to volume; and specific gravity, or the ratio between the mineral's density and that of water. Sometimes minerals can be identified in terms of qualities unique to a specific mineral group or groups: magnetism, radioactivity, fluorescence, phosphorescence, and so on. (For more about mineral characteristics, see Minerals.)

ROCKS. A rock is an aggregate of minerals or organic material, which can appear in consolidated or unconsolidated form. Rocks are of three different types: igneous, formed by crystallization of molten minerals, as in a volcano; sedimentary, usually formed by deposition, compaction, or cementation of weathered rock; and metamorphic, formed by alteration of preexisting rock. Rocks made from organic material are typically sedimentary, an example being coal.

Rocks have possessed economic importance from a time long before "economics" as we know it existed—a time when there was nothing to buy and nothing to sell. That time, of course, would be the Stone Age, which dates back practically to the beginnings of the human species and overlapped with the beginnings of civilization some 5,500 years ago. In the hundreds of thousands of years when stone constituted the most advanced toolmaking material, humans developed an array of stone devices for making fire, sharpening knives, killing animals (and other humans), cutting food or animal skins, and so on.

The Stone Age, both in the popular imagination and (with some qualifications) in actual archaeological fact, was a time when people lived in caves. Since that time, of course, humans have generally departed from the caves, though exceptions exist, as the United States military found in 2001 when attempting to hunt for terrorists in the caves of Afghanistan. In any case, the human attachment to stone dwellings has taken other forms, beginning with the pyramids and continuing through today's masonry homes. Nor is rock simply a structural material for building, as the use of gypsum wallboard, slate countertops, marble finishes, and graveled walkways attests. And, of course, construction is only one of many applications to which rocks and minerals are directed, as we shall see.

Hydrocarbons

As noted earlier, the focus of economic geology is on both rocks and minerals, on the one hand, and fossil fuels, on the other. The latter may be defined as fuel (specifically, coal, oil, and gas) derived from deposits of organic material that have experienced decomposition and chemical alteration under conditions of high pressure. Given this derivation from organic material, by definition all fossil fuels are carbon-based, and, specifically, they are built around hydrocarbons—chemical compounds whose molecules are made up of nothing but carbon and hydrogen atoms.

Theoretically, there is no limit to the number of possible hydrocarbons. Carbon forms itself into apparently limitless molecular shapes, and hydrogen is a particularly versatile chemical partner. Hydrocarbons may form straight chains, branched chains, or rings, and the result is a variety of compounds distinguished not by the elements in their makeup or even (in some cases) by the numbers of different atoms in each molecule, but rather by the structure of a given molecule.

VARIETIES OF HYDROCARBON. Among the various groups of hydrocarbons are alkanes or saturated hydrocarbons, so designated because all the chemical bonds are filled to their capacity (that is, "saturated") with hydrogen atoms. Included among them are such familiar names as methane (CH_4), ethane (C_2H_6), propane (C_3H_8), and butane (C_4H_{10}). The first four, being the lowest in molecular mass, are gases at room temperature, while the heavier ones—including octane (C_8H_{18})—are oily liquids. Alkanes even heavier than octane tend to be waxy solids, an example being paraffin wax, for making candles.

With regard to octane, incidentally, there is a reason why its name is so familiar, while that of heptane (C_7H_{16}) is not. Heptane does not fire smoothly in an internal-combustion engine and therefore disrupts the engine's rhythm. For this reason, it has a rating of zero on a scale of desirability, while octane has a rating of 100. This is why gas stations list octane ratings at the pump: the higher the content of octane, the better the gas is for one's automobile.

In a hydrocarbon chain, if one or more hydrogen atoms is removed, a new bond may be formed. The hydrocarbon chain is then named by adding the suffix *yl*—hence such names as methyl, ethyl, and so on. This indicates that the substance is an alkane, and that something other than hydrogen can be attached to the chain; for example, the attachment of a chlorine atom could yield methyl chloride. Two other large structural groups of hydrocarbons are alkenes and alkynes, which contain double or triple bonds between carbon atoms. Such hydrocarbons are unsaturated—in other words, if the double or triple bond is broken, some of the carbon atoms are then free to form other bonds. Among the products of these groups is the alkene known as acetylene, or C_2H_2, used for welding steel. In addition to alkanes, alkenes, and alkynes, all of which tend to form carbon chains, there are the aromatic hydrocarbons, a traditional name that actually has nothing to do with smell.

All aromatic hydrocarbons contain what is known as a benzene ring, which has the chemical formula C_6H_6 and appears in characteristic ring shapes. In this group are such products as naphthalene, toluene, and dimethyl benzene. These last two are used as solvents as well as in the synthesis of drugs, dyes, and plastics. One of the more famous (or infamous) products in this part of the vast hydrocarbon network is trinitrotoluene, or TNT. Naphthalene is derived from coal tar and used in the synthesis of other compounds. A crystalline solid with a powerful odor, it is found in mothballs and various deodorant disinfectants.

REAL-LIFE APPLICATIONS

Fossil Fuels

The organic material that has decomposed to create the hydrocarbons in fossil fuels comes primarily from dinosaurs and prehistoric plants, though it just as easily could have come from any other organisms that died in large numbers a long, long time ago. To form petroleum, there must be very large quantities of organic material deposited along with sediments and buried under more sediment. The accumulated sediments and organic material are called *source rock*.

What happens after accumulation of this material is critical and depends a great deal on the nature of the source rock. It is important that the organic material—for example, the vast numbers of dinosaurs that died in a mass extinction about 65 million years ago (see Paleontology)—not be allowed simply to rot, as would happen in an aerobic, or oxygen-containing, environment. Instead, the organic material undergoes transformation into hydrocarbons as a result of anaerobic chemical activity, or activity that takes place in the absence of oxygen.

Good source rocks for this transformation are shale or limestone, provided the particular rocks are composed of between 1% and 5% organic carbon. The source rocks should be deep enough that the pressure heats the organic material, yet not so deep that the pressure and temperature cause the rocks to undergo metamorphism or transform them into graphite or other non-hydrocarbon versions of carbon. Temperatures of up to 302°F (150°C) are considered optimal for petroleum generation.

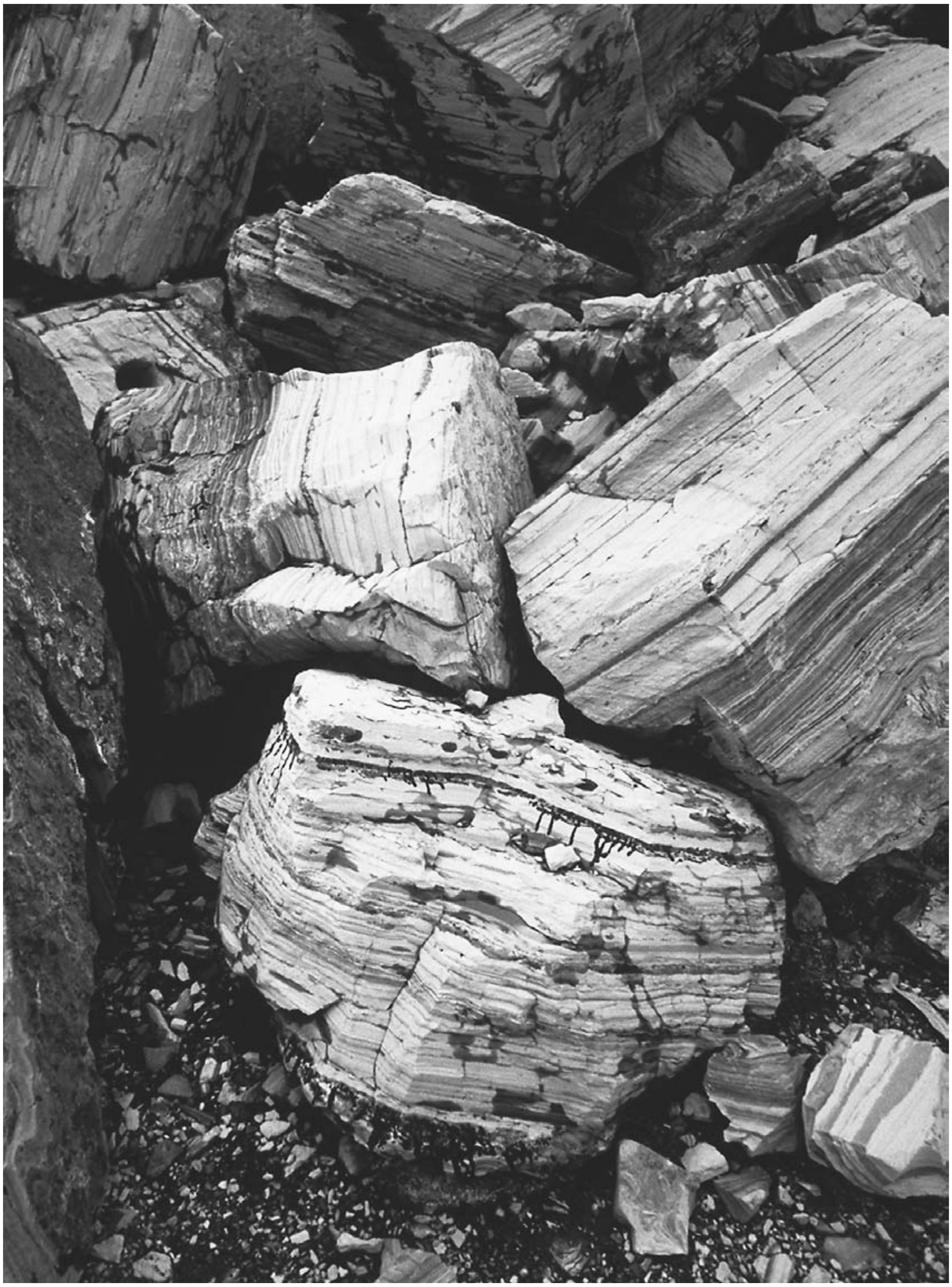
Once generated, petroleum gradually moves from the source rock to a reservoir rock, or a rock that stores petroleum in its pores. A good reservoir rock is one in which the pore space constitutes more than 30% of the rock volume. Yet the rock must be sealed by another rock that is much less porous; indeed, for a seal or cap rock, as it is called, a virtually impermeable rock is preferred. Thus, the best kind of seal-forming rock is one made of very small, closely fitting pieces of sediment, for instance, shale. Such a rock is capable of holding petroleum in place for millions of years until it is ready to be discovered and used.

HUMANS AND PETROLEUM. People have known about petroleum from prehistory, simply because there were places on Earth where it literally seeped from the ground. The modern era of petroleum drilling, however, began in 1853, when an American lawyer named George Bissell (1821–1884) recognized its potential for use as a lamp fuel. He hired "Colonel" Edwin Drake (1819–1880) to oversee the drilling of an oil well at Titusville, Pennsylvania, and in 1859 Drake struck oil. The legend of "black gold," of fortunes to be made by drilling holes in the ground, was born.

In the wake of the development and widespread application of the internal-combustion engine during the latter part of the nineteenth and the early part of the twentieth centuries, interest in oil became much more intense, and wells sprouted up around the world. Sumatra, Indonesia, yielded oil from its first wells in 1885, and in 1901, successful drilling began in Texas—the source of many a Texas-sized fortune. An early form of the company known today as British Petroleum (BP) discovered the first Middle Eastern oil in Persia (now Iran) in 1908. Over the next 50 years, the economic importance and prospects of that region changed considerably.

With the vast expansion in automobile ownership that began following World War I (1914–1918) and reached even greater heights after World War II (1939–1945), the value and importance of petroleum soared. The oil industry boomed, and, as a result, many geologists found employment in a sector that offered far more in the way of financial benefits than university or government positions ever could. Today geologists assist their employers in locating oil reserves, not an easy task because so many variables must line up to produce a viable oil source. Given the cost of drilling a new oil well, which may run to $30 million or more, it is clearly important to exercise good judgment in assessing the possibilities of finding oil.

The oil industry has been fraught with environmental concerns over the impact of drilling (much of which takes place offshore, on rigs placed in the ocean); possible biohazards associated with spills, such as the one involving the Exxon *Valdez* in 1989; and the effect on the



Blocks of shale at the Paraho Oil Shale Facility in Colorado. Shale is known as a seal or cap rock, a virtually impermeable rock made of small, closely fitting pieces of sediment that covers a more porous rock holding petroleum. *(© U.S. Department of Energy/Photo Researchers. Reproduced by permission.)*

atmosphere of carbon monoxide and other greenhouse gases produced by petroleum-burning internal-combustion engines. There is even more wide-ranging concern over United States dependence on oil sources in foreign countries (some of which are openly hostile to the United States) as well as the possible dwindling of resources.

Offshore oil rig at Sabine Pass, Texas. *(© Garry D. McMichael/Photo Researchers. Reproduced by permission.)*

At the present rate of consumption, oil reserves will be exhausted by about the year 2040, but this takes into account only reserves that are considered viable today. As exploration continues, the tapping of United States reserves, such as those in Alaska, will become more and more profitable, leading to increased exploitation of U.S. resources and decreased dependence on oil produced by Middle Eastern states, many of which openly or covertly support terrorist attacks against the United States. In the long run, however, it will be necessary to develop new means of fueling the industrialized world, because petroleum is a nonrenewable resource: there is only so much of it underground, and when it is gone, it will not be replaced for millions of years (if at all).

PETROCHEMICALS. In the meantime, however, petroleum—a mixture of alkanes, alkenes, and aromatic hydrocarbons—makes the world (or at least the industrialized world) go 'round. Petroleum itself is a raw material from which numerous products, collectively known as petrochemicals or petroleum derivatives, are obtained. Through a process termed fractional distillation, the petrochemicals of the lowest molecular mass boil off first, and those having higher mass separate at higher temperatures.

Natural gas separates from petroleum at temperatures below 96.8°F (36°C)—far lower than the boiling point of water. At somewhat higher temperatures, petroleum ether and naphtha, both solvents (naphtha is used in paint thinner), separate; then, in the region between 156.2°F and 165.2°F (69–74°C), gasoline separates. Still higher temperatures yield other substances, each thicker than the one before it: kerosene; fuel for heating and the operation of diesel engines; lubricating oils; petroleum jelly; paraffin wax; and pitch, or tar. A host of other organic chemicals, including various drugs, plastics, paints, adhesives, fibers, detergents, synthetic rubber, and agricultural chemicals, owe their existence to petrochemicals.

Silicon, Silicates, and Other Compounds

It was stated earlier that both carbon and silicon have the tendency to produce long strings of atoms, usually in combination with hydrogen in the first case and oxygen in the second. This is no accident, since silicon lies just below carbon on the periodic table of elements and they share certain chemical features (see Minerals). Just as carbon is at the center of a vast world of hydrocarbons, so silicon is equally important to inorganic substances ranging from sand or silica (SiO_2) to silicone (a highly versatile set of silicon-based products), to the rocks known as silicates.

Silicates are the basis for several well-known mineral types, including garnet, topaz, zircon, kaolinite, talc, mica, and the two most abundant minerals on Earth, feldspar and quartz. (Note that most of the terms used here refer to a group of minerals, not to a single mineral.) Made of compounds formed around silicon and oxygen and comprising various metals, such as aluminum, iron, sodium, and potassium, the silicates account for 30% of all minerals. As such, they appear in everything from gemstones to building materials; yet they are far from the only notable products centered around silicon.

SILICONE AND OTHER COMPOUNDS. Silicone is not a mineral; rather, it is a synthetic product often used as a substitute for organic oils, greases, and rubber. Instead of attaching to oxygen atoms, as in a silicate, silicon atoms in silicone attach to organic groups, that is, molecules containing carbon. Silicone oils frequently are used in place of organic petroleum as a lubricant because they can withstand greater variations in temperature. And because the body tolerates the introduction of silicone implants better than it does organic ones, silicones are used in surgical implants as well. Silicone rubbers appear in everything from bouncing balls to space vehicles, and silicones are also present in electrical insulators, rust preventives, fabric softeners, hair sprays, hand creams, furniture and automobile polishes, paints, adhesives, and even chewing gum.

Even this list does not exhaust the many applications of silicon, which (together with oxygen) accounts for the vast majority of the mass in Earth's crust. Owing to its semimetallic qualities, silicon is used as a semiconductor of electricity. Computer chips are tiny slices of ultrapure silicon, etched with as many as half a million microscopic and intricately connected electronic circuits. These chips manipulate voltages using binary codes, for which 1 means "voltage on" and 0 means "voltage off." By means of these pulses, silicon chips perform multitudes of calculations in seconds—calculations that would take humans hours or months or even years.

A porous form of silica known as silica gel absorbs water vapor from the air and is often packed alongside moisture-sensitive products, such as electronics components, to keep them dry. Silicon carbine, an extremely hard crystalline material manufactured by fusing sand with coke (almost pure carbon) at high temperatures, has applications as an abrasive.

SILICONE, A SYNTHETIC PRODUCT, HAS A VARIETY OF USES, FROM ELECTRICAL INSULATORS TO FABRIC SOFTENERS TO PAINTS AND ADHESIVES. IT IS TOLERATED BETTER BY THE HUMAN BODY THAN ORGANIC COMPOUNDS AND OFTEN IS USED FOR SURGICAL IMPLANTS. *(© Michelle Del Guercio/Photo Researchers. Reproduced by permission.)*

Ores

Earlier, it was stated that an ore is a rock or mineral that possesses economic value. This is true, but a more targeted definition would include the adjective metalliferous, since economically valuable minerals that contain no metals usually are treated as a separate category, industrial minerals. Indeed, it can be said that the interests of economic geology are divided into three areas: ores, industrial minerals, and fuels, which we have discussed already.

The very word ore seems to call to mind one of the oldest-known metals in the world and probably the first material worked by prehistoric metallurgists: gold. Even the Spanish word for gold, *oro,* suggests a connection. When conquistadors from Spain arrived in the New World after about 1500, *oro* was their obsession, and it was said that the Spanish invaders of Mexico found every bit of gold or silver ore located at the surface of the earth. However, miners of the sixteenth century lacked much of the knowledge that helps geologists today find ore deposits that are *not* at the surface.

LOCATING AND EXTRACTING ORES. The modern approach uses knowledge gained from experience. As in Agricola's day, much of the wealth possessed by a mining company is in the form of information regarding the means of best seeking out and retrieving materials from the solid earth. Certain surface geochemical and geophysical indicators help direct the steps of geologists and miners searching for ore. Thus, by the time a company in search of ore begins drilling, a great deal of exploratory work has been done. Only at that point is it possible to determine the value of the deposits, which may simply be minerals of little economic interest.

It is estimated that a cubic mile (1.6 km^3) of average rock contains about $1 trillion worth of metals, which at first sounds promising—until one does the math. A trillion dollars is a lot of money, but 1 cu. mi. (equal to 5,280 $\times$ 5,280 $\times$ 5,280 ft., or 1,609 km^3) is a lot of space too. The result is that 1 cu. ft. (0.028 m^3) is worth only about $6.79. But that is an average cubic foot in an average cubic mile of rock, and no mining company would even consider attempting to extract metals from an average piece of ground. Rather, viable ore appears only in regions that have been subjected to geologic processes that concentrate metals in such a way that their abundance is usually many hundreds of times greater than it would be on Earth as a whole.

Ore contains other minerals, known as *gangue,* which are of no economic value but which serve as a telltale sign that ore is to be found in that region. The presence of quartz, for example, may suggest deposits of gold. Ore may appear in igneous, metamorphic, or sedimentary deposits as well as in hydrothermal fluids. The latter are emanations from igneous rock, in the form of gas or water, that dissolve metals from rocks through which they pass and later deposit the ore in other locations.

CONFRONTING THE HAZARDS OF MINING. Mining, a means of extracting not only ores but many industrial minerals and fuels, such as coal, is difficult work fraught with numerous hazards. There are short-term dangers to the miners, such as cave-ins, flooding, or the release of gases in the mines, as well as long-term dangers that include such mining-related diseases as black lung (typically a hazard of coal miners). Then there is the sheer mental and emotional stress that comes from spending eight or more hours a day away from the sunlight, in claustrophobic surroundings.

And, of course, there is the environmental stress created by mining—not just by the immediate impact of cutting a gash in Earth's surface, which may disrupt ecosystems on the surface, but myriad additional problems, such as the seepage of pollutants into the water table. Abandoned mines present further dangers, including the threat of subsidence, which make these locations unsafe for the long term.

Higher environmental and occupational safety standards, established in the United States during the last third of the twentieth century, have led to changes in the way mining is performed as well as in the way mines are left when the work is completed. For example, mining companies have experimented with the use of chemicals or even bacteria, which can dissolve a metal underground and allow it to be pumped to the surface without the need to create actual underground shafts and tunnels or to send human miners to work them.

Industrial Minerals and Other Products

Industrial minerals, as noted earlier, are nonmetal-containing mineral resources of interest to economic geology. Examples include asbestos, a generic term for a large group of minerals that are highly resistant to heat and flame; boron compounds, which are used for making heat-resistant glass, enamels, and ceramics; phosphates and potassium salts, used in making fertilizers; and sulfur, applied in a range of products, from refrigerants to explosives to purifiers used in the production of sugar.

Just one industrial mineral, corundum (from the oxides class of mineral), can have numerous uses. Extremely hard, corundum in the form of an unconsolidated rock commonly called emery has been used as an abrasive since ancient times. Owing to its very high melting point—even higher than that of iron—corundum also is employed in making alumina, a fire-

KEY TERMS

ALLOY: A mixture of two or more metals.

ATOM: The smallest particle of an element, consisting of protons, neutrons, and electrons. An atom can exist either alone or in combination with other atoms in a molecule.

CHEMICAL BONDING: The joining through electromagnetic force of atoms that sometimes, but not always, represent more than one chemical element.

COMPOUND: A substance made up of atoms of more than one element, chemically bonded to one another.

CONSOLIDATION: A process whereby materials become compacted, or experience an increase in density.

CRYSTALLINE SOLID: A type of solid in which the constituent parts have a simple and definite geometric arrangement that is repeated in all directions.

DEPOSITION: The process whereby sediment is laid down on the Earth's surface.

DUCTILE: Capable of being bent or molded into various shapes without breaking.

ECONOMIC GEOLOGY: The study of fuels, metals, and other materials from the Earth that are of interest to industry or the economy in general.

ELECTRON: A negatively charged particle in an atom, which spins around the nucleus.

FOSSIL FUELS: Fuel derived from deposits of organic material that have experienced decomposition and chemical alteration under conditions of high pressure. These nonrenewable forms of bioenergy include petroleum, coal, peat, natural gas, and their derivatives.

GANGUE: Minerals of no economic value, which appear in nature with ore. Recognition of certain characteristic combinations can help geologists find ore on the basis of its attendant gangue. (The *ue* is silent, as in *tongue.*)

HARDNESS: In mineralogy, the ability of one mineral to scratch another. This can be measured by the Mohs scale.

HYDROCARBON: Any organic chemical compound whose molecules are made up of nothing but carbon and hydrogen atoms.

IGNEOUS ROCK: One of the three principal types of rock, along with sedimentary and metamorphic rock. Igneous rock is formed by the crystallization of molten materials, for instance, in a volcano or other setting where plate tectonic processes take place.

INDUSTRIAL MINERALS: Nonmetallic minerals with uses for industry.

LUSTER: The appearance of a mineral when light reflects off its surface. Among the terms used in identifying luster are metallic, vitreous (glassy), and dull.

METALS: Substances that are ductile, lustrous or shiny in appearance, extremely durable, and excellent conductors of heat and electricity. Metals have very high melting and boiling points, and some (though far from all) have a high degree of hardness.

KEY TERMS CONTINUED

METAMORPHIC ROCK: One of the three principal varieties of rock, along with sedimentary and igneous rock. Metamorphic rock is formed through the alteration of preexisting rock as a result of changes in temperature, pressure, or the activity of fluids. These changes are known as metamorphism.

MINERAL: A naturally occurring, typically inorganic substance with a specific chemical composition and a crystalline structure. Unknown minerals usually can be identified in terms of specific parameters, such as hardness or luster.

MINERALOGY: An area of geology devoted to the study of minerals. Mineralogy includes a number of subdisciplines, such as crystallography, or the study of crystal formations within minerals.

MOHS SCALE: A scale introduced in 1812 by the German mineralogist Friedrich Mohs (1773–1839) that rates the hardness of minerals from 1 to 10. Ten is equivalent to the hardness of a diamond and 1 that of talc, an extremely soft mineral.

MOLECULE: A group of atoms, usually but not always representing more than one element, joined in a structure. Compounds are typically made up of molecules.

NUCLEUS: The center of an atom, a region where protons and neutrons are located and around which electrons spin.

ORE: A metalliferous rock or mineral possessing economic value.

ORGANIC: At one time chemists used the term *organic* only in reference to living things. Now the word is applied to most compounds containing carbon and hydrogen, thus excluding carbonates (which are minerals), and oxides such as carbon dioxide.

PHYSICAL GEOLOGY: The study of the material components of Earth and of the forces that have shaped the planet. Physical geology is one of two principal branches of geology, the other being historical geology.

PROTON: A positively charged particle in an atom.

ROCK: An aggregate of minerals or organic matter, which may be consolidated or unconsolidated.

SEDIMENT: Material deposited at or near Earth's surface from a number of sources, most notably preexisting rock.

SEDIMENTARY ROCK: One of the three major types of rock, along with igneous and metamorphic rock. Sedimentary rock usually is formed by the deposition, compaction, and cementation of rock that has experienced weathering. It also may be formed as a result of chemical precipitation.

STREAK: The color of the powder produced when one mineral is scratched by another, harder one.

UNCONSOLIDATED ROCK: Rock that appears in the form of loose particles, such as sand.

proof product used in furnaces and fireplaces. Though pure corundum is colorless, trace amounts of certain elements can yield brilliant colors: hence, corundum with traces of chromium becomes a red ruby, while traces of iron, titanium, and other elements yield varieties of sapphire in yellow, green, and violet as well as the familiar blue.

AN ARRAY OF APPLICATIONS. We have only begun to scratch the surface, as it were, of the uses to which minerals can be put: after all, everything—literally, every solid object—that people use is either organic in origin or a mineral. The wide array of applications of minerals is clear from the following list of mineral categories, classified by application: abrasives (corundum, diamond), ceramics (feldspar, quartz), chemical minerals (halite, sulfur, borax), and natural pigments (hematite, limonite).

Lime, cement, and plaster comes from calcite and gypsum, while building materials—both structural and ornamental—are products of agate, as well as the two aforementioned minerals. Table salt is a mineral, and so is chalk, as are countless other products. There are rocks, such as granite and marble, used in building, decoration, or artwork, and then there are "rocks"—to use a word that is at once a geologic term and a slang expression—that appear in the form of jewelry.

JEWELRY. Out of all minerals, 16 are important for their use as gems: beryl, chrysoberyl, corundum, diamond, feldspar, garnet, jade, lazurite, olivine, opal, quartz, spinel, topaz, tourmaline, turquoise, and zircon. Not all forms of these minerals, of course, are precious. Furthermore, some minerals provide more than one type of gem: corundum, as we have noted, is a source of rubies and sapphires, while beryl produces both emeralds and aquamarines.

Note that many of the precious gems familiar to most of us are not minerals in their own right but versions of minerals. At least one, the pearl, is not on this list because, with its organic origin, it is not a mineral. Certainly not all minerals are created equal: even in the list of 16 just provided, the name *diamond* stands out, representing a worldwide standard of value. Yet a diamond is nothing but pure carbon, which also appears in the form of graphite and (with a very few impurities) as coke for burning.

A diamond is unusual, however, in many respects, including the fact that it is basically a huge "molecule" composed of carbon atoms strung together by chemical bonds. The size of this formation corresponds to the size of the diamond, such that a diamond of 1 carat is simply a gargantuan "molecule" containing about 10^{22} (10,000,000,000,000,000,000,000, or 10 billion trillion) carbon atoms. Not only is a diamond rare and (when properly selected, cut, and polished) extremely beautiful, it is also extraordinarily hard. At the top of the Mohs scale, it can cut any other substance, but nothing can cut a diamond except another diamond.

WHERE TO LEARN MORE

Atlas of Rocks, Minerals, and Textures (Web site). <http://www.geosci.unc.edu/Petunia/IgMetAtlas/mainmenu.html>.

Bates, Robert Latimer. *Industrial Minerals: How They Are Found and Used.* Hillside, NJ: Enslow Publishers, 1988.

McGraw-Hill Encyclopedia of Science and Technology. 8th ed. New York: McGraw-Hill, 1997.

The Mineral and Gemstone Kingdom: Minerals A–Z (Web site). <http://www.minerals.net/mineral/>.

"Minerals and Metals: A World to Discover." Natural Resources Canada (Web site). <http://www.nrcan.gc.ca/mms/school/e_mine.htm>.

Ohio Department of Natural Resources Division of Mineral Resources Management (Web site). <http://www.dnr.state.oh.us/mineral/index.html>.

Spitz, Peter H. *Petrochemicals: The Rise of an Industry.* New York: John Wiley and Sons, 1988.

Stevens, Paul. *Oil and Gas Dictionary.* New York: Nichols, 1988.

Symes, R. F. *Rocks and Minerals.* Illus. Colin Keates and Andreas Einsiedel. New York: Dorling Kindersley, 2000.

Western Australia Department of Mineral and Petroleum Resources (Web site). <http://www.dme.wa.gov.au/index.html>.

GEOPHYSICS

GRAVITY AND GEODESY

GEOMAGNETISM

CONVECTION

ENERGY AND EARTH

GRAVITY AND GEODESY

CONCEPT

Thanks to the force known as gravity, Earth maintains its position in orbit around the Sun, and the Moon in orbit around Earth. Likewise, everything on and around Earth holds its place—the waters of the ocean, the gases of the atmosphere, and so on—owing to gravity, which is also the force that imparts to Earth its nearly spherical shape. Though no one can really say what gravity *is,* it can be quantified in terms of mass and the inverse of the distance between objects. Earth scientists working in the realm of geophysics known as geodesy measure gravitational fields, as well as anomalies within them, for a number of purposes, ranging from the prediction of tectonic processes to the location of oil reserves.

HOW IT WORKS

Gravitation

Not only does gravity keep Earth and all other planets in orbit around the Sun, it also makes it possible for our solar system to maintain its position in the Milky Way, rather than floating off through space. Likewise, the position of our galaxy within the larger universe is maintained because of gravity. As for the universe itself, though many questions remain about its size, mass, and boundaries, it seems clear that the cosmos is held together by gravity.

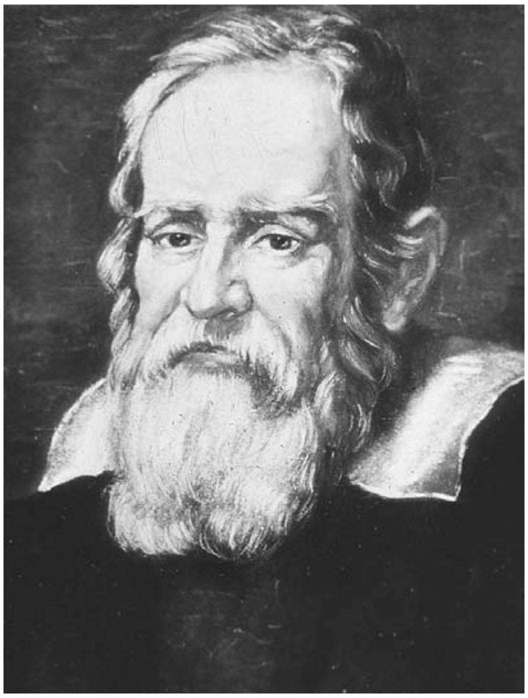
Thanks to gravity, all objects on Earth as well as those within its gravitational field remain fixed in place. These objects include man-made satellites, which have grown to number in the thousands since the first was launched in 1957, as well as the greatest satellite of them all: the Moon. Even though people are accustomed to thinking of gravity in these large terms, with regard to vast bodies such as Earth or the Moon, every object in the universe, in fact, exerts some gravitational pull on another.

This attraction is proportional to the product of the mass of the two bodies, and inversely related to the distance between them. Bodies have to be fairly large (i.e., larger than an asteroid) for this attraction to be appreciable, but it is there, and thus gravity acts as a sort of "glue" holding together the universe. As to what gravity really is or exactly why it works, both of which are legitimate questions, the answers have so far largely eluded scientists.

Present-day scientists are able to understand how gravity works, however, inasmuch as it can be described as a function of mass and the gravitational constant (discussed later) and an inverse function of distance. They also are able to measure gravitational fields and anomalies within them. That, in fact, is the focus of geodesy, an area of geophysics devoted to the measurement of Earth's shape and gravitational field.

Discovering Gravity

As discussed in several places within this book (see the entries Earth, Science, and Nonscience and Studying Earth), the physical sciences made little progress until the early sixteenth century. For centuries, the writings of the Greek philosopher Aristotle (384–322 B.C.) and the Alexandrian astronomer Ptolemy (*ca.* A.D. 100–170) had remained dominant, reinforcing an almost entirely erroneous view of the universe. This Aristotelian/Ptolemaic universe had Earth at its



GALILEO GALILEI *(Corbis-Bettmann. Reproduced by permission.)*

center, with the Sun, Moon, and other planets orbiting it in perfect circles.

The discovery by the Polish astronomer Nicolaus Copernicus (1473–1543) that Earth rotates on its axis and revolves around the Sun ultimately led to the overturning of the Ptolemaic model. This breakthrough, which inaugurated the Scientific Revolution (*ca.* 1550–1700), opened the way for the birth of physics, chemistry, and geology as genuine sciences. Copernicus himself was a precursor to this revolution rather than its leader; by contrast, the Italian astronomer Galileo Galilei (1564–1642) introduced the principles of study, known as the scientific method, that govern the work of scientists to this day.

GALILEO AND GRAVITATIONAL ACCELERATION. Galileo applied his scientific method in his studies of falling objects and was able to show that objects fall as they do, not because of their weight (as Aristotle had claimed) but as a consequence of gravitational force. This meant that the acceleration of all falling bodies would have to be the same, regardless of weight.

Of course, everyone knows that a stone falls faster than a feather, but Galileo reasoned that this was a result of factors other than weight, and later investigations confirmed that air resistance, rather than weight, is responsible for this difference. In other words, a stone falls faster than a feather not because it is heavier but because the feather encounters greater air resistance. In a vacuum, or an area devoid of all matter, including air, they would fall at the same rate.

On the other hand, if one drops two objects that meet similar air resistance but differ in weight—say, a large stone and a smaller one—they fall at almost exactly the same rate. To test this hypothesis directly, however, would have been difficult for Galileo: stones fall so fast that even if dropped from a great height, they would hit the ground too soon for their rate of fall to be tested with the instruments then available.

Instead, Galileo used the motion of a pendulum and the behavior of objects rolling or sliding down inclined planes as his models. On the basis of his observations, he concluded that all bodies are subject to a uniform rate of gravitational acceleration, later calibrated at 32 ft. (9.8 m) per second per second. What this means is that for every 32 ft. an object falls, it is accelerating at a rate of 32 ft. per second as well; hence, after two seconds it falls at the rate of 64 ft. per second, after three seconds it falls at 96 ft. per second, and so on.

NEWTON'S BREAKTHROUGH. Building on the work of his distinguished forebear, Sir Isaac Newton (1642–1727), who was born the same year Galileo died, developed a paradigm for gravitation that even today explains the behavior of objects in virtually all situations throughout the universe. Indeed, the Newtonian model reigned supreme until the early twentieth century, when Albert Einstein (1879–1955) challenged it on certain specifics.

Even so, Einstein's relativity did not *disprove* the Newtonian system as Copernicus and Galileo had disproved Aristotle's and Ptolemy's theories; rather, it showed the limitations of Newtonian mechanics for describing the behavior of certain objects and phenomena. In the ordinary world of day-to-day experience, however, the Newtonian system still offers the key to how and why things work as they do. This is particularly the case with regard to gravity.

UNDERSTANDING THE LAW OF UNIVERSAL GRAVITATION

Like Galileo, Newton began in part with the aim of testing hypotheses put forward by an astronomer—in this case, Johannes Kepler (1571–1630). In the early years of the seventeenth century, Kepler published his three laws of planetary motion, which together identified the elliptical (oval-shaped) path of the planets around the Sun. Kepler had discovered a mathematical relationship that connected the distances of the planets from the Sun to the period of their revolution around it. Like Galileo with Copernicus, Newton sought to generalize these principles to explain not only *how* the planets moved but also *why* they did so.

The result was Newton's *Philosophiae naturalis principia mathematica* (Mathematical principles of natural philosophy, 1687). Usually referred to simply as the *Principia,* the book proved to be one of the most influential works ever written. In it, Newton presented his law of universal gravitation, along with his three laws of motion. These principles offered a new model for understanding the mechanics of the universe.

THE THREE LAWS OF MOTION

Newton's three laws of motion may be summarized in this way:

- An object at rest will remain at rest, and an object in motion will remain in motion at a constant velocity unless and until outside forces act upon it.
- The net force acting upon an object is a product of its mass multiplied by its acceleration.
- When one object exerts a force on another, the second object exerts on the first a force equal in magnitude but opposite in direction.

The first law of motion identifies inertia, a concept introduced by Galileo to explain what kept the planets moving around the Sun. Inertia is the tendency of an object either to keep moving or to keep standing still, depending on what it is already doing. Note that the first law refers to an object moving at a constant velocity: velocity is speed in a certain direction, so a constant velocity would be the same speed in the same direction.

Inertia is measured by mass, which—as the second law states—is a component of force and is inversely related to acceleration. The latter, as defined by physics, has a much broader meaning than it usually is given in ordinary life. Acceleration does not mean simply an increase of speed for an object moving in a straight line; rather, it is a change in velocity—that is, a change of speed or direction or both.

By definition, then, rotational motion (such as that of Earth around the Sun) involves acceleration, because any movement other than motion in a straight line at a constant speed requires a change in velocity. This further means that an object experiencing rotational motion *must* be under the influence of some force. That force is gravity, and as the third law shows, every force exerted in one direction is matched by an equal force in the opposing direction. (This law is sometimes rendered "For every action there is an equal and opposite reaction.")

NEWTON'S GRAVITATIONAL FORMULA. The law of universal gravitation can be stated as a formula for calculating the gravitational attraction between two objects of a certain mass, m_1 and m_2: $F_{grav} = G \times (m_1 m_2)/r^2$. In this equation, F_{grav} is gravitational force, and r^2 is the square of the distance between the two objects.

As for *G,* in Newton's time the value of this number was unknown. Newton was aware simply that it represented a very small quantity: without it, $(m_1 m_2)/r^2$ could be quite sizable for objects of relatively great mass separated by a relatively small distance. When multiplied by this very small number, however, the gravitational attraction would be revealed to be very small as well. Only in 1798, more than a century after Newton's writing, did the English physicist Henry Cavendish (1731–1810) calculate the value of *G* using a precision instrument called a torsion balance.

The value of *G* is expressed in units of force multiplied by distance squared, and then divided by,mass squared; in other words, *G* is a certain value of $(N \times m^2)/kg^2$, where *N* stands for newtons, *m* for meters, and *kg* for kilograms. Nor is the numerical value of *G* a whole number such as 1. A figure as large as 1, in fact, is astronomically huge compared with *G,* whose value is 6.67×10^{-11}—in other words, 0.0000000000667.

PHYSICAL GEODESY

Within the realm of geodesy is that of physical geodesy, which is concerned specifically with the measurement of Earth's gravitational field as well as the geoid. The latter may be defined as a surface of uniform gravitational potential covering the entire Earth at a height equal to sea level. ("Potential" here is analogous to height or, more specifically, position in a field. For a discussion of potential in a gravitational field, see Energy and Earth.)

Thus, in areas that are above sea level, the geoid would be below ground—indeed, *far* below it in mountainous regions. Yet in some places (most notably the Dead Sea and its shores, the lowest point on Earth), it would be above the solid earth and waters. The geoid is also subject to deviations or anomalies, owing to the fact that the planet's mass is not distributed uniformly; in addition, small temporary disturbances in the geoid may occur on the seas as a result of wind, tides, and currents.

Generally speaking, however, the geoid is a stable reference platform from which to measure gravitational anomalies. It is a sort of imaginary gravitational "skin" covering the planet, and in the past, countries conducting geodetic surveys tended to choose a spot within their boundaries as the reference point for all measurements. With the development of satellites and their use for geodetic research, however, it has become more common for national geodetic societies to use global points of reference such as the planet's center of mass.

MEASUREMENTS FROM SPACE, LAND, AND SEA. The geoid can be determined by using such a satellite, equipped with a radar altimeter, but there is also the much older technique of terrestrial gravity measurement. The terrestrial method is much more difficult and prone to error, however, and calculations require detailed checking and correction to remove potential anomalies due to the presence of matter in areas above the points at which gravitational measurements were obtained.

Also highly subject to error are measurements made from a vessel at sea. This has to do not only with the effect of the ship's pitch and roll but also with something called the Eötvös effect. Named for the Hungarian physicist Baron Roland Eötvös (1848–1919), who conducted extensive studies on gravity, the effect is related to the Coriolis force, which causes the deflection of atmospheric and oceanic currents in response to Earth's rotation. Measurements of gravity from the air are also subject to the Eötvös effect, though the use of GPS (global positioning system) information, obtained from satellites, can improve greatly the accuracy of seaborne measurements.

HOW GRAVITY IS MEASURED. Scientists can obtain absolute terrestrial gravity measurements by measuring the amount of time it takes for a pellet to fall a certain distance within a vacuum—that is, a chamber from which all matter, including air, has been removed. This, of course, is the same technology Galileo used in making his observations more than 400 years ago. It is also possible to obtain relative gravity measurements with the use of mechanical balance instruments.

As noted earlier, the acceleration due to gravity is 9.8 m/s^2, or 9.8 m s^{-2}. (Scientists sometimes use the latter notation, in which the minus sign is not meant to indicate a negative but rather is used in place of "per".) This number is the measure of Earth's gravitational field. In measuring gravitational anomalies, scientists may use the Gal, named after Galileo, which is equal to 0.01 m/s^2. Typically, however, the milligal, equal to one-thousandth of a Gal, is used. Note that "Gal" sometimes is rendered in lowercase, but this can be confusing, because it looks like the abbreviation for "gallon. ")

WHY MEASURE GRAVITY? Why is it important to measure gravity and gravitational anomalies? One answer is that weight values can vary considerably, depending on one's position relative to Earth's gravitational field. A fairly heavy person might weigh as much as a pound less at the equator than at the poles and less still at the top of a high mountain. The value of the gravity field at sea level has a range from 9.78 to 9.83 m/s^2, a difference of about 50,000 g.u., and it is likely to be much lower than 9.78 m/s^2 at higher altitudes.

Indeed, the higher one goes, the weaker Earth's gravitational field becomes. At the same time, the gases of the atmosphere dissipate, which is the reason why it is hard to breathe on high mountains without an artificial air supply and impossible to do so in the stratosphere or above it. At the upper edge of the mesosphere, Earth's gravitational field is no longer strong enough to hold large quantities of hydrogen, lightest of all elements, which constitutes the atmosphere at that point. Beyond the mesosphere, the atmosphere simply fades away, because there is not sufficient gravitational force to hold its particles in place.

Back down on Earth, gravity measurements are of great importance to the petroleum industry, which uses them to locate oil-containing salt domes. Furthermore, geologists, in general, remain acutely interested in measurements of gravity, the force behind tectonics, or the deformation of Earth's crust. Thus, gravity, responsible for fashioning Earth's exterior into the nearly spherical shape it has, is key to the shaping of its interior as well.

REAL-LIFE APPLICATIONS

GRAVITY ON EARTH

Using Newton's gravitational formula, it is relatively easy to calculate the pull of gravity between two objects. It is also easy to see why the attraction is insignificant unless at least one of the objects has enormous mass. In addition, application of the formula makes it clear why G is such a tiny number.

Suppose two people each have a mass of 45.5 kg—equal to 100 lb. on Earth, though not on the Moon, a matter that will be explained later in this essay—and they stand 1 m (3.28 ft.) apart. Thus, m_1m_2 is equal to 2,070 kg (4,555 lb.), and r^2 is equal to 1 m squared. Applied to the gravitational formula, this figure is rendered as 2,070 $kg^2/1\ m^2$. This number then is multiplied by the gravitational constant, and the result is a net gravitational force of 0.000000138 N (0.00000003 lb.)—about the weight of a single-cell organism!

WEIGHT. What about Earth's gravitational force on one of those people? To calculate this force, we could apply the formula for universal gravitation, substituting Earth for m_2, especially because the mass of Earth is known: 5.98×10^{24} kg, or 5.98 septillion (1 followed by 24 zeroes) kg. We know the value of that mass, in fact, through the application of Newton's laws and the formulas derived from them. But for measuring the gravitational force between something as massive as Earth and something as small as a human body, it makes more sense to apply instead the formula embodied in Newton's second law of motion: $F = ma$. (Force equals mass multiplied by acceleration.)

For a body of any mass on Earth, acceleration is figured in terms of g—the acceleration due to gravity, which, as noted earlier, is equal to 32 ft. (9.8 m) per second squared. (Note, also, that this is a lowercase g, as opposed to the uppercase G that represents the gravitational constant.) Using the metric system, by multiplying the appropriate mass figure in kilograms by 9.8 m/s^2, one would obtain a value in newtons (N). To perform the same calculation with the English system, used in America, it would be necessary first to calculate the value of mass in slugs (which, needless to say, is a little-known unit) and multiply it by 32 $ft./s^2$ to yield a value in pounds.

In both cases, the value obtained, whether in newtons or pounds, is a measure of weight rather than of mass, which is measured in kilograms or slugs. For this reason, it is not entirely accurate to say that 1 kg is equal to 2.2 lb. This is true on Earth, but it would not be true on the Moon. The kilogram is a unit of mass, and as such it would not change anywhere in the universe, whereas the pound is a unit of force (in this case, gravitational force) and therefore varies according to the rate of acceleration for the gravitational field in which it is measured. For this reason, scientists prefer to use figures for mass, which is one of the fundamental properties (along with length, time, and electric charge) of the universe.

WHY EARTH IS ROUND—AND NOT ROUND

Everyone knows that Earth, the Sun, and all other large bodies in space are "round" (i.e., spherical), but why is that true? The reason is that gravity will not allow them to be otherwise: for any large object, the gravitational pull of its interior forces the surface to assume a relatively uniform shape. The most uniform of three-dimensional shape is that of a sphere, and the larger the mass of an object, the greater its tendency toward sphericity.

Earth has a relatively small vertical differential between its highest and lowest surface points, Mount Everest (29,028 ft., or 8,848 m) on the Nepal-Tibet border and the Mariana Trench (–36,198 ft., –10,911 m) in the Pacific Ocean, respectively. The difference is just 12.28 mi. (19.6 km)—not a great distance, considering that Earth's radius is about 4,000 mi. (6,400 km).

On the other hand, an object of less mass is more likely to retain a shape that is far less than spherical. This can be shown by reference to the Martian moons Phobos and Deimos, both of which are oblong—and both of which are tiny, in

THE FORCE OF GRAVITY IMPARTS A SPHERICAL SHAPE TO EARTH BECAUSE OF ITS LARGE MASS. AN OBJECT OF LESS MASS WILL HAVE A FAR LESS SPHERICAL SHAPE. THE MARTIAN MOON PHOBOS, SHOWN HERE, IS OBLONG OWING TO ITS TINY MASS. *(© Julian Baum/Photo Researchers. Reproduced by permission.)*

terms of size and mass, compared with Earth's Moon. Mars itself has a radius half that of Earth, yet its mass is only about 10% of Earth's. In light of what has been said about mass, shape, and gravity, it should not be surprising to learn that Mars is also home to the tallest mountain in the solar system, the volcano Olympus Mons, which stands 16 mi. (27 km) high.

EARTH'S 'FLAT TOP' (AND BOTTOM). With regard to gravitation, a spherical object behaves as though its mass were concentrated near its center, and indeed, 33% of Earth's mass is at its core, even though the core accounts for only about 20% of the planet's volume. Geologists believe that the composition of Earth's core must be molten iron, which creates the planet's vast electromagnetic field.

It should be noted, however, that Earth is not really a perfect sphere, and the idea that its mass is concentrated at its center, while it works well in general, poses some problems in making exact gravitational measurements. If Earth were standing still, it would be much nearer to the shape of a sphere; however, it is not standing still but instead rotates on its axis, as does every other object of any significance in the solar system.

Incidentally, if Earth were suddenly to stop spinning, the gases in the atmosphere would keep moving at their current rate of about 1,000 MPH (1,600 km/h). They would sweep over the planet with the force of the greatest hurricane ever known, ripping up everything but the mountains. As to why Earth spins at all, scientists are not entirely sure. It may well be angular momentum (the momentum associated with rotational motion) imparted to it at some point in the very distant past, perhaps because it and the rest of the solar system were once part of a vast spinning cloud.

At any rate, the fact that Earth is spinning on its axis creates a certain centripetal, or inward-pulling, force, and this force produces a corresponding centrifugal (outward) component. To understand this concept, consider what happens to a sample of blood when it is rotated in a centrifuge. When the centrifuge spins, centripetal force pulls the material in the vial toward the center of the spin, but the material with greater mass has more inertia and therefore responds less to centripetal force. As a result, the heavier red blood cells tend to stay at the bottom of the vial (or, as it is spinning, on the outside), while the lighter plasma is pulled inward. The result is the separation between plasma and red blood cells.

Where Earth is concerned, this centrifugal component of centripetal force manifests as an equatorial bulging. Simply put, Earth's diameter around the equator is greater than at the poles, which are slightly flattened. The difference is small—the equatorial diameter of Earth is about 26.72 mi. (43 km) greater than the polar diameter—but it is not insignificant. In fact, as noted later, a person of fairly significant weight actually would notice a difference if he or she got on the scales at the equator (say, in Singapore) and then later weighed in near one of the poles (for instance, in the Norwegian possession of Svalbard, the northernmost human settlement on Earth).

Owing to this departure from a perfectly spherical shape, the Sun and Moon exert additional torques on Earth, and these torques cause shifts in the position of the planet's rotational axis in space. An imaginary line projected from

KEY TERMS

ACCELERATION: A change in velocity over time. The acceleration due to gravity, for instance, is 32 ft. (9.8 m) per second per second, meaning that for every second an object falls, its velocity is increasing as well.

ATMOSPHERE: In general, an atmosphere is a blanket of gases surrounding a planet. Unless otherwise identified, however, the term refers to the atmosphere of Earth, which consists of nitrogen (78%), oxygen (21%), argon (0.93%), and other substances that include water vapor, carbon dioxide, ozone, and noble gases such as neon, which together comprise 0.07%.

CENTRIFUGAL: A term describing the tendency of objects in uniform circular motion to move outward, away from the center of the circle. Though the term centrifugal force often is used, it is inertia, rather than force, that causes the object to move outward.

CENTRIPETAL FORCE: The force that causes an object in uniform circular motion to move inward, toward the center of the circle.

FORCE: The product of mass multiplied by acceleration.

GEODESY: An area of geophysics devoted to the measurement of Earth's shape and gravitational field.

GEOID: A surface of uniform gravitational potential covering the entire Earth at a height equal to sea level.

GEOPHYSICS: A branch of the earth sciences that combines aspects of geology and physics. Geophysics addresses the planet's physical processes as well as its gravitational, magnetic, and electric properties and the means by which energy is transmitted through its interior.

INERTIA: The tendency of an object at rest to remain at rest or an object in motion to remain in motion, at a uniform velocity, at a uniform velocity, unless acted upon by some outside force.

MASS: A measure of inertia, indicating the resistance of an object to a change in its motion.

POTENTIAL: Position in a field, such as a gravitational force field.

SCIENTIFIC METHOD: A set of principles and procedures for systematic study that includes observation; the formation of hypotheses, theories, and laws; and continual testing and reexamination.

SCIENTIFIC REVOLUTION: A period of accelerated scientific discovery that completely reshaped the world. Usually dated from about 1550 to 1700, the Scientific Revolution saw the origination of the scientific method and the introduction of ideas such as the heliocentric (Sun-centered) universe and gravity.

TORQUE: A force that produces, or tends to produce, rotational motion.

UNIFORM CIRCULAR MOTION: The motion of an object around the center of a circle in such a manner that speed is constant or unchanging.

VACUUM: An area devoid of matter, even air.

VELOCITY: Speed in a certain direction.

WEIGHT: A measure of the gravitational force on an object. Weight thus would change from planet to planet, whereas mass remains constant throughout the universe. A pound is a unit of weight, whereas a kilogram is a unit of mass.

the North Pole and into space therefore, over a period of time, would appear to move. In the course of about 25,800 years, this point (known as the celestial north pole) describes the shape of a cone, a movement known as Earth's *precession.*

Satellites

Why, then, does Earth move around the Sun, or the Moon around Earth? As should be clear from Newton's gravitational formula and the third law of motion, the force of gravity works both ways: not only does a stone fall toward Earth, but Earth also actually falls *toward it.* The mass of Earth is so great compared with that of the stone that the movement of Earth is imperceptible—but it does happen.

Furthermore, because Earth is round, when one hurls a projectile at a great distance, Earth curves away from the projectile. Eventually, gravity itself forces the projectile to the ground. If one were to fire a rocket at 17,700 mi. per hour (28,500 km per hour), however, something unusual would happen. At every instant of time, the projectile would be falling toward Earth with the force of gravity—*but* the curved Earth would be falling away from it at the same rate. Hence, the projectile would remain in constant motion around the planet—that is, it would be in orbit.

The same is true of an artificial satellite's orbit around Earth: even as the satellite falls toward Earth, Earth falls away from it. Change the names of the players, and this same relationship exists between Earth and its great natural satellite, the Moon. Furthermore, it is the same with the Sun and *its* many satellites, including Earth: Earth plunges toward the Sun with every instant of its movement, but at every instant, the Sun falls away.

Where to Learn More

Ardley, Neil. *The Science Book of Gravity.* San Diego, CA: Harcourt Brace Jovanovich, 1992.

Ask the Space Scientist (Web site). <http://image.gsfc.nasa.gov/poetry/ask/askmag.html>.

Beiser, Arthur. *Physics,* 5th ed. Reading, MA: Addison-Wesley, 1991.

Exploring Gravity (Web site). <http://www.curtin.edu.au/curtin/dept/phys-sci/gravity/>.

Geodesy for the Layman (Web site). <http://www.nima.mil/GandG/geolay/TR80003A.HTM>.

The Gravity Society (Web site). <http://www.gravity.org>.

Hancock, Paul L., and Brian J. Skinner. *The Oxford Companion to the Earth.* New York: Oxford University Press, 2000.

Riley, Peter D. *Earth.* Des Plaines, IL: Heinemann Interactive Library, 1998.

Smith, David G. *The Cambridge Encyclopedia of Earth Sciences.* New York: Cambridge University Press, 1981.

Wilford, John Noble. *The Mapmakers.* New York: Knopf, 2000.

GEOMAGNETISM

CONCEPT

Scientists have long recognized a connection between electricity and magnetism, but the specifics of this connection, along with the recognition that electromagnetism is one of the fundamental interactions in the universe, were worked out only in the mid–nineteenth century. By that time, geologists had come to an understanding of Earth as a giant magnet. This was the principle that made possible the operation of compasses, which greatly aided mariners in navigating the seas: magnetic materials, it so happened, point northward. As it turns out, however, Earth's magnetic North Pole is not the same as its geographic one, and even the pole's northerly location is not a permanent fact. Once upon a time and, in fact, at many times in Earth's history, the magnetic North Pole lay at the southern end of the planet.

HOW IT WORKS

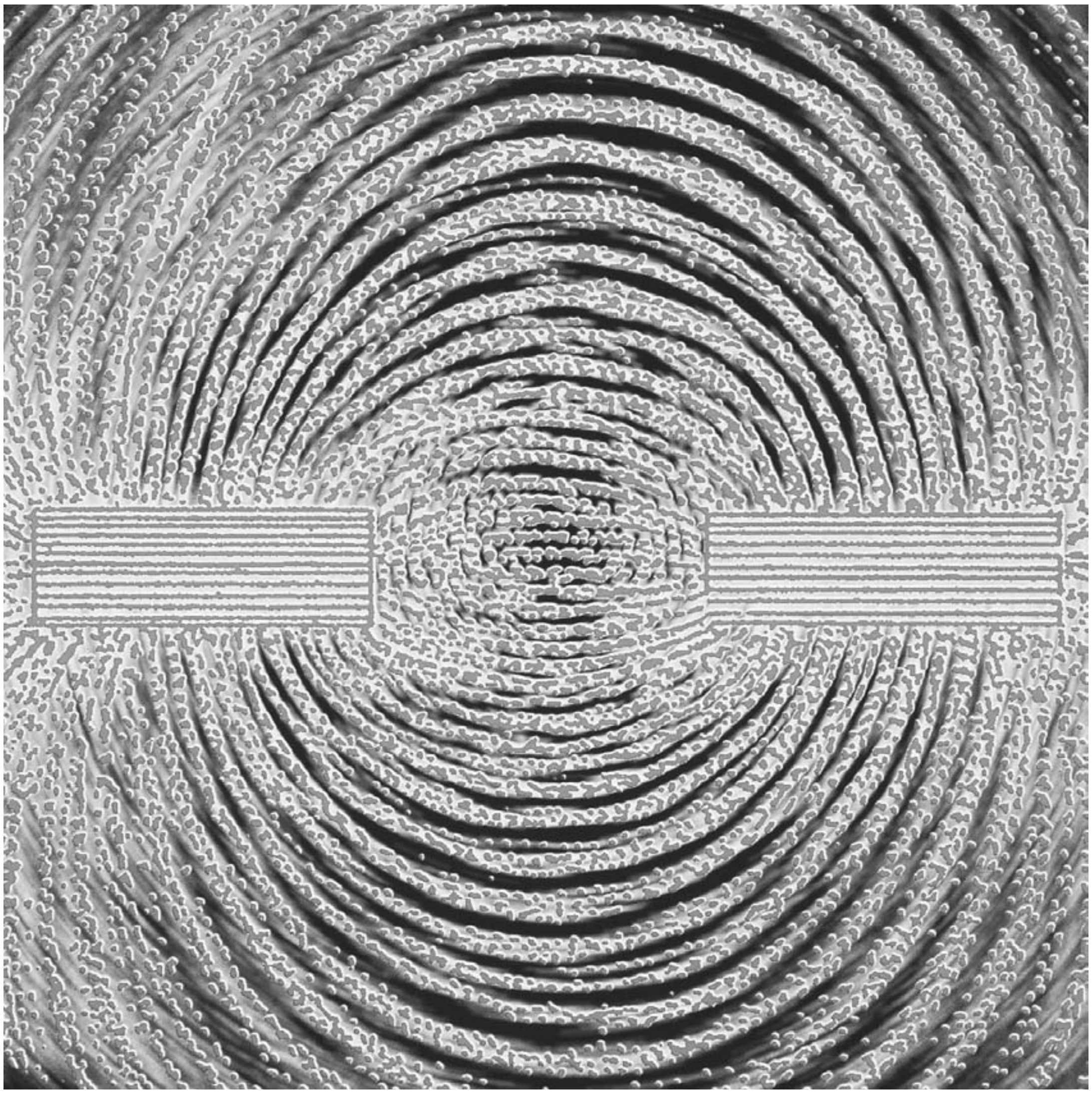
Electromagnetism

The Greek philosopher Thales (640?–546 B.C.) was the first to observe that when amber is rubbed with certain types of materials, the friction imparts to it the ability to pick up light objects. The word electricity comes from the Greek word for amber, *elektron,* and, in fact, magnetism and electricity are simply manifestations of the same force. This concept of electric and magnetic interaction seems to have been established early in human history, though it would be almost 2,500 years before scientists came to a mature understanding of the relationship.

As in so much else, studies in electromagnetism made little progress from the time of the Romans to the late Renaissance, a span of nearly 1,500 years. Yet it is worth noting that thefirst ideas scientists had about studying Earth's history scientifically came from observing the planet's magnetic field. In the course of his work on that subject, the English astronomer Henry Gellibrand (1597–1636) showed that the field has changed over time. This suggested that it would be possible to form hypotheses about the planet's past, even though humans had no direct information regarding the origins of Earth. Thus, Gellibrand (who, ironically, was also a minister) helped make it possible for geologists to move beyond a strict interpretation of the Bible in studying the history of Earth. (See Earth, Science, and Nonscience for more on the Genesis account and its interpretations.)

ELECTROMAGNETIC STUDIES COME OF AGE. Beginning in the 1700s, a number of thinkers conducted experiments concerning the nature of electricity and magnetism and the relationship between them. Among these thinkers were several giants in physics and other disciplines, including one of America's greatest founding fathers, Benjamin Franklin (1706–1790). In addition to his famous (and highly dangerous) experiment with lightning, Franklin also contributed the names positive and negative to the differing electric charges discovered earlier by the French physicist Charles Du Fay (1698–1739).

In 1785 the French physicist and inventor Charles Coulomb (1736–1806) established the basic laws of electrostatics and magnetism. He maintained that there is an attractive force that, like gravitation, can be explained in terms of the



The magnetic field around two bar magnets. *(© A. Pasieka/Photo Researchers. Reproduced by permission.)*

inverse of the square of the distance between objects (see Gravity and Geodesy). That attraction itself, however, results not from gravity but from electric charge, according to Coulomb.

A few years later, the German mathematician Karl Friedrich Gauss (1777–1855) developed a mathematical theory for finding the magnetic potential of any point on Earth. His contemporary, the Danish physicist Hans Christian Oersted (1777–1851), became the first scientist to establish a clear relationship between electricity and magnetism. This led to the formalization of electromagnetism, the branch of physics devoted to the study of electric and magnetic phenomena.

The French mathematician and physicist André Marie Ampère (1775–1836) concluded that magnetism is the result of electricity in motion, and in 1831 the British physicist and chemist Michael Faraday (1791–1867) published his theory of electromagnetic induction. This theory shows how an electric current in one coil can set up a current in another through the development of a magnetic field, and it enabled Faraday's development of the first generator. For the first time in history humans were able to convert mechanical energy systematically into electric energy.

ELECTROMAGNETIC FORCE. By this point scientists were convinced that a relationship between electricity and magnetism existed, yet they did not know exactly *how* the two related. Then, in 1865, the Scottish physicist James Clerk Maxwell (1831–1879) published a groundbreaking paper, "On Faraday's Lines of Force," in which he outlined a theory of electromagnetic force. The latter may be defined as the total force on an electrically charged particle, which is a combination of forces due to electric and magnetic fields around the particle.

Maxwell thus had discovered a type of force in addition to gravity, and this reflected a "new" type of fundamental interaction, or a basic mode by which particles interact in nature. Nearly two centuries earlier Sir Isaac Newton (1642–1727) had identified the first, gravity, and in the twentieth century two other forms of fundamental interaction—strong nuclear and weak nuclear—were identified as well.

In his work Maxwell drew on the studies conducted by his predecessors but added a new statement. According to Maxwell, electric charge is conserved, meaning that the sum total of electric charge in the universe does not change, though it may be redistributed. This statement, which did not contradict any of the experimental work done by other physicists, was based on Maxwell's predictions regarding what should happen in situations of electromagnetism. Subsequent studies have supported his predictions.

Magnetism

What, then, is the difference between electricity and magnetism? It is primarily a matter of orientation. When two electric charges are at rest, it appears to an observer that the force between them is merely electric. If the charges are in motion relative to the observer, it appears as though a different sort of force, known as magnetism, exists between them.

An electromagnetic wave, such as that which is emitted by the Sun, carries both an electric and a magnetic component at mutually perpendicular angles. If you extend your hand, palm flat, with the fingers straight and the thumb pointing at a 90° angle to the fingers, the direction that the fingers are pointing would be that of the electromagnetic wave. Your thumb points in the direction of the electric field, and the flat of your palm indicates the direction of the magnetic field, which is perpendicular both to the electric field and to the direction of wave propagation.

A field, in this sense, is a region of space in which it is possible to define the physical properties of each point in the region at any given moment in time. Thus, an electric field and a magnetic field are simply regions in which the electric and magnetic components, respectively, of electromagnetic force are exerted.

MAGNETISM AT THE ATOMIC LEVEL. At the atomic level magnetism is the result of motion by electrons (negatively charged subatomic particles) in relation to one another. Rather like planets in a solar system, electrons revolve around the atom's nucleus and rotate on their own axes. (In fact, the exact nature of their movement is much more complex, but this analogy is accurate enough for the present purposes.) Both types of movement create a magnetic force field between electrons, and as a result the electron takes on the properties of a tiny bar magnet with a north pole and south pole. Surrounding this infinitesimal magnet are lines of magnetic force, which begin at the north pole and curve outward, describing an ellipse as they return to the south pole.

In most atoms, electrons are paired such that their magnetic fields cancel out one another. However, in certain cases, such as when there is an odd electron or when other factors become more significant, the fields line up to create what is known as a net magnetic dipole, or a unity of direction. These elements, among them, iron, cobalt, and nickel as well as various alloys or mixtures, are commonly known as magnetic metals, or natural magnets.

MAGNETIZATION. Magnetization occurs when an object is placed in a magnetic field. In this field magnetic force acts on a moving charged particle such that the particle would experience no force if it moved in the direction of the magnetic field. In other words, it would be "drawn," as a ten-penny nail is drawn to a common bar or horseshoe (U-shaped) magnet. An electric current is an example of a moving charge, and, indeed, one of the best ways to create a magnetic field is with a current. Often this is done by means of a solenoid, a current-carrying wire coil through which the material to be magnetized is passed, much as one would pass a straight wire up through the interior of a spring.

When a natural magnet becomes magnetized (that is, when a magnetic metal or alloy comes into contact with an external magnetic field), a change occurs at the level of the domain, a group of atoms equal in size to about 5×10^{-5} meters across—just large enough to be visible under a microscope.

In an unmagnetized sample, there may be an alignment of unpaired electron spins within a domain, but the direction of the various domains' magnetic forces in relation to one another is random. Once a natural magnet is placed within an external magnetic force field, however, one of two things happens to the domains. Either they all come into alignment with the field or, in certain types of material, those domains in alignment with the field grow, while the others shrink to nonexistence.

The first of these processes is called domain alignment, or ferromagnetism, and the second is termed domain growth, or ferrimagnetism. Both processes turn a natural magnet into what is known as a permanent magnet—or, more simply, a magnet. The magnet is then capable of temporarily magnetizing a ferromagnetic item, as, for instance, when one rubs a paper clip against a permanent magnet and then uses the magnetized clip to lift other paper clips. Of the two varieties, however, a ferromagnet is stronger, because it requires a more powerful magnetic force field to become magnetized. Most powerful of all is a saturated ferromagnetic metal, one in which all the unpaired electron spins are aligned.

REAL-LIFE APPLICATIONS

The Magnetic Compass

A bar magnet placed in a magnetic field will rotate until it lines up with the field's direction. The same thing happens when one suspends a magnet from a string: it lines up with Earth's magnetic field and points in a north-south direction. The Chinese of the first century B.C. discovered that a strip of magnetic metal always tends to point toward geographic north, though they were unaware of the electromagnetic force that causes this to happen.

This led ultimately to the development of the magnetic compass, which typically consists of a magnetized iron needle suspended over a card marked with the four cardinal directions (north, south, east, and west). The needle is attached to a pivoting mechanism at its center, which allows it to move freely so that the tip of the needle will always point the user northward. The magnetic compass proved so important that it typically is ranked alongside paper, printing, and gunpowder as one of premodern China's four great gifts to the West. Before the compass, mariners had to depend purely on the position of the Sun and other, less reliable means of determining direction; hence, the invention quite literally helped open up the world.

ODD BEHAVIOR OF THE COMPASS. The compass, in fact, helped make possible the historic voyage of Christopher Columbus (1451–1506) in 1492. While sailing across the Atlantic, Columbus noticed something odd: his compass did not always point toward what he knew, based on the Sun's position, to be geographic north. The further he traveled, the more he noticed this phenomenon, which came to be known as magnetic declination.

When Columbus returned to Europe and reported on his observations of magnetic declination (along with the much bigger news of his landing in the New World, which he thought was Asia), his story perplexed mariners. Eventually, European scientists worked out tables of magnetic declination, showing the amount of deviation at various points on Earth, and this seemed to allay sailors' concerns.

Then, in 1544, the German astronomer Georg Hartmann (1489–1564) observed that a freely floating magnetized needle did not always stay perfectly horizontal and actually dipped more and more strongly as he traveled north. When he was moving south, on the other hand, the needle tended to become more closely horizontal. For many years, this phenomenon, along with magnetic declination, remained perplexing. Nor, for that matter, did scientists understand exactly how or why a compass works. Then, in 1600, the English physicist William Gilbert (1540–1603) became the first to suggest a reason.

Earth as a Giant Magnet

Gilbert coined the terms electric attraction, electric force, and magnetic pole. In De magnete (On the magnet), he became the first thinker to introduce the idea, now commonly accepted, that Earth itself is a giant magnet. Not only does it have north and south magnetic poles, but it also is surrounded by vast arcs of magnetic force, called the geomagnetic field. (The term geomagnetism, as opposed to magnetism, refers to the magnetic properties of Earth as a whole rather than those possessed by a single object or place on Earth.)

In the paragraphs that follow, we discuss the shape of this magnetic field, including the positions of the magnetic north and south poles; the origins of the field, primarily in terms of the known or suspected physical forces that sustain it (as opposed to the original cause of Earth's magnetic field, a more complicated and speculative subject); as well as changes in the magnetic field. Those changes, along with techniques for measuring the geomagnetic field, also are discussed at other places in this book.

EARTH'S MAGNETIC FIELD. Hartmann's compass phenomenon can be explained by the fact that Earth is a magnet and that its north and south magnetic poles are close to the geographic north and south. As for the phenomenon observed by Columbus, it is a result of the difference between magnetic and geographic north. If one continued to follow a compass northward, it would lead not to Earth's North Pole but to a point identified in 1984 as 77°N, 102°18' W—that is, in the Queen Elizabeth Islands of far northern Canada.

Earlier we described the lines that make up the magnetic force field around a bar magnet. A field of similar shape, though, of course, of much larger size (yet still invisible), also surrounds Earth. From the magnetic north and south poles, lines of magnetic force rise into space and form giant curves that come back around and reenter Earth at the opposing pole, so that the planet is surrounded by a vast series of concentric loops. If one could draw a straight line through the center of all these loops, it would reach Earth at a point 11° from the equator. Likewise the north and south magnetic poles—which are on a plane perpendicular to that of Earth's magnetic field—are 11° off the planet's axis.

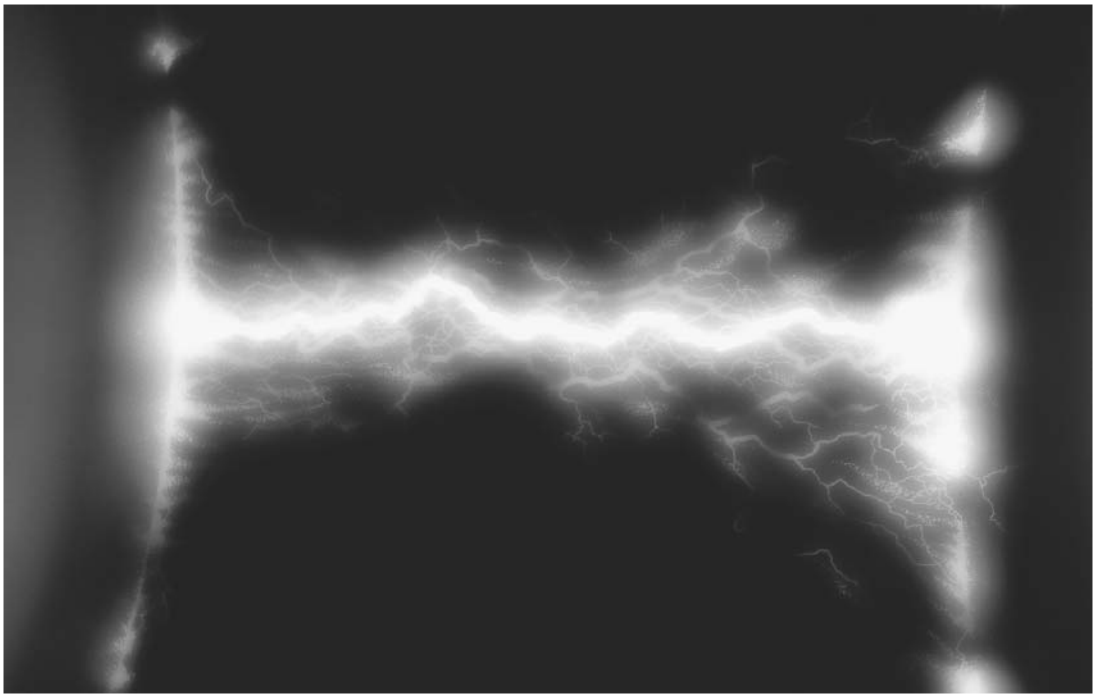
THE MAGNETOSPHERE. Surrounding the planet is a vast region called the magnetosphere, an area in which ionized particles (i.e., ones that have lost or gained electrons so as to acquire a net electric charge) are affected by Earth's magnetic field. The magnetosphere is formed by the interaction between our planet's magnetic field and the solar wind, a stream of particles from the Sun. (See Sun, Moon, and Earth for more about the solar wind.) Its shape would be akin to that of Earth's magnetic field, as described earlier, were it not for the Sun's influence.

The side of the magnetosphere closer to the Sun does indeed resemble the giant series of concentric loops described earlier. These loops are enormous, such that the forward, or sunlit, edge of the magnetosphere is located at a distance of some 10 Earth radii (about 35,000 mi., or 65,000 km). On the other side from the Sun—the rear, or dark, side—the shape of the magnetosphere is quite different. Instead of forming relatively small loops that curve right back around into Earth's poles, the lines of magnetic force on this side shoot straight out into space a distance of some 40 Earth radii (about 140,000 mi., or 260,000 km).

The shape of the magnetosphere, then, is a bit like that of a comet moving toward the Sun. Surrounding it is the magnetopause, a sort of magnetic dead zone about 62 mi. (100 km) thick, which shields Earth from most of the solar wind. In front of it (toward the Sun) is an area of magnetic turbulence known as the magnetosheath, and still closer to the Sun is a boundary called the bow shock, a shock-wave front that slows particles of solar wind considerably. Since Earth is turning, the side of the planet away from the Sun is continually changing, of course, but the shape of the magnetosphere remains more or less intact. It is, however, highly affected by solar activity, such that an increase in solar wind can cause a depression in the magnetosphere. (See Sun, Moon, and Earth for a discussion of auroras, produced by an interaction between the solar wind and the magnetosphere.)

The Source of Earth's Magnetic Field

As noted earlier, scientists at present understand little with regard to the *origins* of Earth's magnetic field—that is, the original action or actions that resulted in the creation of a geomagnetic field that has remained active for billions of years. No less a figure than Albert Einstein (1879–1955) identified this question as one of the great unsolved problems of science. On the other hand, scientists do have a good understanding of the geomagnetic field's *source,* in terms of the physical conditions that make it possible. (This distinction is rather like that between efficient cause and material cause, as discussed in Earth, Science, and Nonscience.)



The electric discharge between two metal objects. *(© P. Jude/Photo Researchers. Reproduced by permission.)*

It is believed that the source of Earth's magnetism lies in a core of molten iron some 4,320 mi. (6,940 km) across, constituting half the planet's diameter. Within this core run powerful electric currents that create the geomagnetic field. Actually, the field seems to originate in the outer core, consisting of an iron-nickel alloy that is kept fluid owing to the exceedingly great temperatures. The materials of the outer core undergo convection, vertical circulation that results from variations in density brought about by differences in temperature.

This process of convection (imagine giant spirals moving vertically through the molten metal) creates the equivalent of a solenoid, described earlier. Even so, there had to be an original source for the magnetic field, and it is possible that it came from the Sun. In any case, the magnetic field could not continue to exist if the fluid of the outer core were not in constant convective motion. If this convection stopped, within about 10,000 years (which, in terms of Earth's life span, is like a few seconds to a human being), the geomagnetic field would decay and cease to operate. Likewise, if Earth's core ever cooled and solidified, Earth would become like the Moon, a body whose magnetic field has disappeared, leaving only the faintest traces of magnetism in its rocks.

Changes in the Magnetic Field

Though there is no reason to believe that anything so dramatic will happen, there are curious and perhaps troubling signs that Earth's magnetic field is changing. According to data recorded by the U.S. Geological Survey, which updates information on magnetic declination, the field is shifting—and weakening. Over the course of about a century, scientists have recorded data suggesting a reduction of about 6% in the strength of the magnetic field.

The behavior, in terms of both weakening and movement, appears to be similar to changes taking place in the magnetic field of the Sun. Indeed, as we have seen already, Earth's magnetism is heavily affected by the Sun, and it is possible that a period of strong solar-flare activity could shut down Earth's magnetic field. Even the present trend of weakening, if it were to continue for just 1,500 years, would wipe out the magnetic field. Some scientists believe that the planet is simply experiencing a fluctuation, however, and that the geomagnetic field will recover. Others

KEY TERMS

CONSERVATION: In physics and other sciences, "to conserve" something means "to result in no net loss of" that particular component. It is possible that within a given system the component may change form or position, but as long as the net value of the component remains the same, it has been conserved.

DIPOLE: A pair of equal and opposite electric charges, or an entire body having the characteristics of a dipole—for instance, a magnet with north and south poles.

ELECTROMAGNETIC ENERGY: A form of energy with electric and magnetic components that travels in waves.

ELECTROMAGNETIC FORCE: The total force on an electrically charged particle, which is a combination of forces due to electric and magnetic fields around the particle. Electromagnetic force reflects electromagnetic interaction, one of the four fundamental interactions in nature.

ELECTRON: A negatively charged particle in an atom, which spins around the nucleus.

ELEMENT: A substance made up of only one kind of atom. Unlike compounds, elements cannot be broken chemically into other substances.

FIELD: A region of space in which it is possible to define the physical properties of each point in the region at any given moment in time.

FUNDAMENTAL INTERACTION: The basic mode by which particles interact. There are four known fundamental interactions in nature: gravitational, electromagnetic, strong nuclear, and weak nuclear.

GEOMAGNETISM: A term referring to the magnetic properties of Earth as a whole rather than those possessed by a single object or place on Earth.

MAGNETIC DECLINATION: The angle between magnetic north and geographic north.

MAGNETOSPHERE: An area surrounding Earth, reaching far beyond the atmosphere, in which ionized particles (i.e., ones that have lost or gained electrons so as to acquire a net electric charge) are affected by Earth's magnetic field.

PALEOMAGNETISM: An area of historical geology devoted to studying the direction and intensity of magnetic fields in the past, as discerned from the residual magnetization of rocks.

POTENTIAL: Position in a field, such as a gravitational force field.

SOLAR WIND: A stream of particles continually emanating from the Sun and moving outward through the solar system.

SYSTEM: Any set of interactions that can be set apart mentally from the rest of the universe for the purposes of study, observation, and measurement.

maintain that the geomagnetic field is on its way to a reversal.

A reversal? Odd as it may sound, the direction of the geomagnetic field has reversed itself many, many times in the past. Furthermore, the planet has attempted unsuccessfully to reverse its geomagnetic field many more times—as recently as 30,000 to 40,000 years ago. These reversals are among the interests of paleomagnetism, the area of geology devoted to the direction and intensity of magnetic fields in the past, as discerned from the residual magnetization of rocks.

A compass works, of course, because the metal points toward Earth's magnetic north pole, which is close to its geographic north pole. Likewise, the magnetic materials in the rocks of Earth point north—or rather, they *would* point north if the direction of the magnetic field had not changed over time. Around the turn of the nineteenth century, geologists noticed that whereas some magnetic rocks pointed toward Earth's current North Pole, some were pointing in the opposite direction. This led to the realization that the magnetic field had reversed and to the development of paleomagnetism as a field of study. Studies in paleomagnetism, in turn, have provided confirmation of the powerful theory known as plate tectonics. (See Plate Tectonics for more on paleomagnetism and the shifting of plates beneath Earth's surface.)

WHERE TO LEARN MORE

Campbell, Wallace Hall. *Earth Magnetism: A Guided Tour Through Magnetic Fields.* New York: Harcourt Academic Press, 2001.

Dispezio, Michael A., and Catherine Leary. *Awesome Experiments in Electricity and Magnetism.* New York: Sterling, 1998.

Geomagnetism (Web site). <http://ub.nmh.ac.uk/>.

The Great Magnet, the Earth (Web site). <http://istp.gsfc.nasa.gov/earthmag/dmglist.htm>.

Hancock, Paul L., and Brian J. Skinner. *The Oxford Companion to the Earth.* New York: Oxford University Press, 2000.

Laboratory of Earth's Magnetism (Web site). <http://geo.phys.spbu.ru/Magnetism/>.

Lafferty, Peter. *Magnets to Generators.* Illus. Alex Pang and Galina Zolfanghari. New York: Gloucester Press, 1989.

Lefkowitz, R. J., and J. Carol Yake. *Forces in the Earth: A Book About Gravity and Magnetism.* New York: Parents' Magazine Press, 1976.

Smith, David G. *The Cambridge Encyclopedia of Earth Sciences.* New York: Cambridge University Press, 1981.

Stern, David P. *400 Years of "De magnete"* (Web site). <http://www-istp.gsfc.nasa.gov/earthmag/demagint.htm>.

CONVECTION

CONCEPT

Convection is the name for a means of heat transfer, as distinguished from conduction and radiation. It is also a term that describes processes affecting the atmosphere, waters, and solid earth. In the atmosphere, hot air rises on convection currents, circulating and creating clouds and winds. Likewise, convection in the hydrosphere circulates water, keeping the temperature gradients of the oceans stable. The term convection generally refers to the movement of fluids, meaning liquids and gases, but in the earth sciences, convection also can be used to describe processes that occur in the solid earth. This geologic convection, as it is known, drives the plate movement that is one of the key aspects of plate tectonics.

HOW IT WORKS

Introduction to Convection

Some concepts and phenomena cross disciplinary boundaries within the earth sciences, an example being the physical process of convection. It is of equal relevance to scientists working in the geologic, atmospheric, and hydrologic sciences, or the realms of study concerned with the geosphere, atmosphere, and hydrosphere, respectively. The only major component of the earth system not directly affected by convection is the biosphere, but given the high degree of interconnection between different subsystems, convection indirectly affects the biosphere in the air, waters, and solid earth.

Convection can be defined as vertical circulation that results from differences in density ultimately brought about by differences in temperature, and it involves the transfer of heat through the motion of hot fluid from one place to another. In the physical sciences, the term fluid refers to any substance that flows and therefore has no definite shape. This usually means liquids and gases, but in the earth sciences it can refer even to slow-flowing solids. Over the great expanses of time studied by earth scientists, the net flow of solids in certain circumstances (for example, ice in glaciers) can be substantial.

Convection and Heat

As indicated in the preceding paragraph, convection is related closely to heat and temperature and indirectly related to another phenomenon, thermal energy. What people normally call *heat* is actually thermal energy, or kinetic energy (the energy associated with movement) produced by molecules in motion relative to one another.

Heat, in its scientific meaning, is internal thermal energy that flows from one body of matter to another or from a system at a higher temperature to a system at a lower temperature. Temperature thus can be defined as a measure of the average molecular kinetic energy of a system. Temperature also governs the direction of internal energy flow between two systems. Two systems at the same temperature are said to be in a state of thermal equilibrium; when this occurs, there is no exchange of heat, and therefore heat exists only in transfer between two systems.

There is no such thing as cold, only the absence of heat. If heat exists only in transit between systems, it follows that the direction of heat flow must *always* be from a system at a higher temperature to a system at a lower temperature. (This fact is embodied in the second law of thermodynamics, which is discussed, along with other topics mentioned here, in Energy and Earth.) Heat transfer occurs through three means: conduction, convection, and radiation.

CONDUCTION AND RADIATION. Conduction involves successive molecular collisions and the transfer of heat between two bodies in contact. It usually occurs in a solid. Convection requires the motion of fluid from one place to another, and, as we have noted, it can take place in a liquid, a gas, or a near solid that behaves like a slow-flowing fluid. Finally, radiation involves electromagnetic waves and requires no physical medium, such as water or air, for the transfer.

If you put one end of a metal rod in a fire and then touch the "cool" end a few minutes later, you will find that it is no longer cool. This is an example of heating by conduction, whereby kinetic energy is passed from molecule to molecule in the same way as a secret is passed from one person to another along a line of people standing shoulder to shoulder. Just as the original phrasing of the secret becomes garbled, some kinetic energy is inevitably lost in the series of transfers, which is why the end of the rod outside the fire is still much cooler than the one sitting in the flames.

As for radiation, it is distinguished from conduction and convection by virtue of the fact that it requires no medium for its transfer. This explains why space is cold yet the Sun's rays warm Earth: the rays are a form of electromagnetic energy, and they travel by means of radiation through space. Space, of course, is the virtual absence of a medium, but upon entering Earth's atmosphere, the heat from the electromagnetic rays is transferred to various media in the atmosphere, hydrosphere, geosphere, and biosphere. That heat then is transferred by means of convection and conduction.

HEAT TRANSFER THROUGH CONVECTION. Like conduction and unlike radiation, convection requires a medium. However, in conduction the heat is transferred from one molecule to another, whereas in convection the heated fluid itself is actually moving. As it does, it removes or displaces cold air in its path. The flow of heated fluid in this situation is called a convection current.

Convection is of two types: natural and forced. Heated air rising is an example of natural convection. Hot air has a lower density than that of the cooler air in the atmosphere above it and therefore is buoyant; as it rises, however, it loses energy and cools. This cooled air, now denser than the air around it, sinks again, creating a repeating cycle that generates wind.

Forced convection occurs when a pump or other mechanism moves the heated fluid. Examples of forced-convection apparatuses include some types of ovens and even refrigerators or air conditioners. As noted earlier, it is possible to transfer heat only from a high-temperature reservoir to a low-temperature one, and thus these cooling machines work by removing hot air. The refrigerator pulls heat from its compartment and expels it to the surrounding room, while an air conditioner pulls heat from a room or building and releases it to the outside.

Forced convection does not necessarily involve man-made machines: the human heart is a pump, and blood carries excess heat generated by the body to the skin. The heat passes through the skin by means of conduction, and at the surface of the skin it is removed from the body in a number of ways, primarily by the cooling evaporation of perspiration.

REAL-LIFE APPLICATIONS

Convective Cells

One important mechanism of convection, whether in the air, water, or even the solid earth, is the convective cell, sometimes known as the convection cell. The latter may be defined as the circular pattern created by the rising of warmed fluid and the sinking of cooled fluid. Convective cells may be only a few millimeters across, or they may be larger than Earth itself.

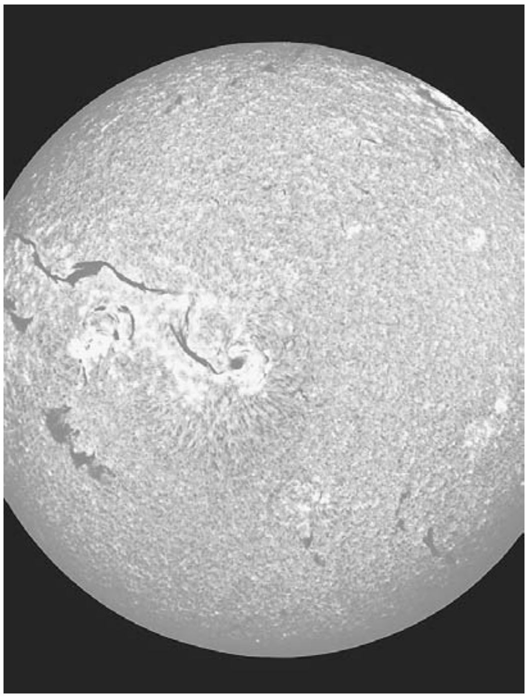
These cells can be observed on a number of scales. Inside a bowl of soup, heated fluid rises, and cooled fluid drops. These processes are usually hard to see unless the dish in question happens to be one such as Japanese miso soup. In this case, pieces of soybean paste, or miso, can be observed as they rise when heated and then drop down into the interior to be heated again.

On a vastly greater scale, convective cells are present in the Sun. These vast cells appear on the Sun's surface as a grainy pattern formed by the

A CUMULONIMBUS CLOUD—THUNDERHEAD—IS A DRAMATIC EXAMPLE OF A CONVECTION CELL. *(© Keith Kent/Photo Researchers. Reproduced by permission.)*

variations in temperature between the parts of the cell. The bright spots are the top of rising convection currents, while the dark areas are cooled gas on its way to the solar interior, where it will be heated and rise again.

A cumulonimbus cloud, or "thunderhead," is a particularly dramatic example of a convection cell. These are some of the most striking cloud formations one ever sees, and for this reason the director Akira Kurosawa used scenes of



CONVECTIVE CELLS APPEAR ON THE SUN'S SURFACE AS A GRAINY PATTERN FORMED BY VARIATIONS IN TEMPERATURE. *(© Noao/Photo Researchers. Reproduced by permission.)*

rolling thunderheads to add an atmospheric quality (quite literally) to his 1985 epic *Ran.* In the course of just a few minutes, these vertical towers of cloud form as warmed, moist air rises, then cools and falls. The result is a cloud that seems to embody both power and restlessness, hence Kurosawa's use of cumulonimbus clouds in a scene that takes place on the eve of a battle.

A SEA BREEZE. Convective cells, along with convection currents, help explain why there is usually a breeze at the beach. At the seaside, of course, there is a land surface and a water surface, both exposed to the Sun's light. Under such exposure, the temperature of land rises more quickly than that of water. The reason is that water has an extraordinarily high specific heat capacity—that is, the amount of heat that must be added to or removed from a unit of mass for a given substance to change its temperature by 33.8°F (1°C). Thus a lake, stream, or ocean is always a good place to cool down on a hot summer day.

The land, then, tends to heat up more quickly, as does the air above it. This heated air rises in a convection current, but as it rises and thus overcomes the pull of gravity, it expends energy and therefore begins to cool. The cooled air then sinks. And so it goes, with the heated air rising and the cooling air sinking, forming a convective cell that continually circulates air, creating a breeze.

CONVECTIVE CELLS UNDER OUR FEET. Convective cells also can exist in the solid earth, where they cause the plates (movable segments) of the lithosphere—the upper layer of Earth's interior, including the crust and the brittle portion at the top of the mantle—to shift. They thus play a role in plate tectonics, one of the most important areas of study in the earth sciences. Plate tectonics explains a variety of phenomena, ranging from continental drift to earthquakes and volcanoes. (See Plate Tectonics for much more on this subject.)

Whereas the Sun's electromagnetic energy is the source of heat behind atmospheric convection, the energy that drives geologic convection is geothermal, rising up from Earth's core as a result of radioactive decay. (See Energy and Earth.) The convective cells form in the asthenosphere, a region of extremely high pressure at a depth of about 60–215 mi. (about 100–350 km), where rocks are deformed by enormous stresses.

In the asthenosphere, heated material rises in a convection current until it hits the bottom of the lithosphere (the upper layer of Earth's interior, comprising the crust and the top of the mantle), beyond which it cannot rise. Therefore it begins moving laterally or horizontally, and as it does so, it drags part of the lithosphere. At the same time, this heated material pushes away cooler, denser material in its path. The cooler material sinks lower into the mantle (the thick, dense layer of rock, approximately 1,429 mi. [2,300 km] thick, between Earth's crust and core) until it heats again and ultimately rises up, thus propagating the cycle.

SUBSIDENCE: FAIR WEATHER AND FOUL

As with convective cells, subsidence can occur in the atmosphere or geosphere. The term subsidence can refer either to the process of subsiding, on the part of air or solid earth, or, in the case of solid earth, to the resulting formation. It thus is defined variously as the downward movement of air, the sinking of ground, or a depression in the earth. In the present context we will discuss atmospheric subsidence, which is more closely related to convection. (For more about geologic

KEY TERMS

ASTHENOSPHERE: A region of extremely high pressure underlying the lithosphere, where rocks are deformed by enormous stresses. The asthenosphere lies at a depth of about 60–215 mi. (about 100–350 km).

ATMOSPHERE: In general, an atmosphere is a blanket of gases surrounding a planet. Unless otherwise identified, however, the term refers to the atmosphere of Earth, which consists of nitrogen (78%), oxygen (21%), argon (0.93%), and other substances that include water vapor, carbon dioxide, ozone, and noble gases such as neon, which together comprise 0.07%.

BIOSPHERE: A combination of all living things on Earth—plants, animals, birds, marine life, insects, viruses, single-cell organisms, and so on—as well as all formerly living things that have not yet decomposed.

CONDUCTION: The transfer of heat by successive molecular collisions. Conduction is the principal means of heat transfer in solids, particularly metals.

CONVECTION: Vertical circulation that results from differences in density ultimately brought about by differences in temperature. Convection involves the transfer of heat through the motion of hot fluid from one place to another and is of two types, natural and forced. (See *natural convection, forced convection.*)

CONVECTION CURRENT: The flow of material heated by means of convection.

CONVECTIVE CELL: The circular pattern created by the rising of warmed fluid and the sinking of cooled fluid. This is sometimes called a convection cell.

CORE: The center of Earth, an area constituting about 16% of the planet's volume and 32% of its mass. Made primarily of iron and another, lighter element (possibly sulfur), it is divided between a solid inner core with a radius of about 760 mi. (1,220 km) and a liquid outer core about 1,750 mi. (2,820 km) thick.

CRUST: The uppermost division of the solid earth, representing less than 1% of its volume and varying in depth from 3 to 37 mi. (5 to 60 km). Below the crust is the mantle.

FLUID: In the physical sciences, the term fluid refers to any substance that flows and therefore has no definite shape. Fluids can be both liquids and gases. In the earth sciences, occasionally substances that appear to be solid (for example, ice in glaciers) are, in fact, flowing slowly.

FORCED CONVECTION: Convection that results from the action of a pump or other mechanism (whether man-made or natural), directing heated fluid toward a particular destination.

GEOSPHERE: The upper part of Earth's continental crust, or that portion of the solid earth on which human beings live and which provides them with most of their food and natural resources.

HEAT: Internal thermal energy that flows from one body of matter to another.

HYDROSPHERE: The entirety of Earth's water, excluding water vapor in the atmosphere but including all oceans, lakes, streams, groundwater, snow, and ice.

KINETIC ENERGY: The energy that an object possesses by virtue of its motion.

KEY TERMS CONTINUED

LITHOSPHERE: The upper layer of Earth's interior, including the crust and the brittle portion at the top of the mantle.

MANTLE: The dense layer of rock, approximately 1,429 mi. (2,300 km) thick, between Earth's crust and its core.

NATURAL CONVECTION: Convection that results from the buoyancy of heated fluid, which causes it to rise.

PLATE TECTONICS: The name both of a theory and of a specialization of tectonics. As an area of study, plate tectonics deals with the large features of the lithosphere and the forces that shape them. As a theory, it explains the processes that have shaped Earth in terms of plates and their movement.

PLATES: Large, movable segments of the lithosphere.

RADIATION: The transfer of energy by means of electromagnetic waves, which require no physical medium (for example, water or air) for the transfer. Earth receives the Sun's energy via the electromagnetic spectrum by means of radiation.

SUBSIDENCE: A term that refers either to the process of subsiding, on the part of air or solid Earth, or, in the case of solid Earth, to the resulting formation. Subsidence thus is defined variously as the downward movement of air, the sinking of ground, or a depression in Earth's crust.

SYSTEM: Any set of interactions that can be set apart mentally from the rest of the universe for the purposes of study, observation, and measurement.

TECTONICS: The study of tectonism, including its causes and effects, most notably mountain building.

TECTONISM: The deformation of the lithosphere.

TEMPERATURE: The direction of internal energy flow between two systems when heat is being transferred. Temperature measures the average molecular kinetic energy in transit between those systems.

THERMAL ENERGY: Heat energy, a form of kinetic energy produced by the motion of atomic or molecular particles in relation to one another. The greater the relative motion of these particles, the greater the thermal energy.

subsidence, see the entries Geomorphology and Mass Wasting.)

In the atmosphere, subsidence results from a disturbance in the normal upward flow of convection currents. These currents may act to set up a convective cell, as we have seen, resulting in the flow of breeze. The water vapor in the air may condense as it cools, changing state to a liquid and forming clouds. Convection can create an area of low pressure, accompanied by converging winds, near Earth's surface, a phenomenon known as a cyclone. On the other hand, if subsidence occurs, it results in the creation of a high-pressure area known as an anticyclone.

Air parcels continue to rise in convective currents until the density of their upper portion is equal to that of the surrounding atmosphere, at which point the column of air stabilizes. On the other hand, subsidence may occur if air at an altitude of several thousand feet becomes denser than the surrounding air without necessarily being cooler or moister. In fact, this air is unusually dry, and it may be warm or cold. Its density then makes it sink, and, as it does, it compresses the air around it. The result is high pressure at the surface and diverging winds just above the surface.

The form of atmospheric subsidence described here produces pleasant results, explaining why high-pressure systems usually are associated with fair weather. On the other hand, if the subsiding air settles onto a cooler lay of air, it creates what is known as a subsidence inversion, and the results are much less beneficial. In this situation a warm air layer becomes trapped between cooler layers above and below it, at a height of several hundred or even several thousand feet. This means that air pollution is trapped as well, creating a potential health hazard. Subsidence inversions occur most often in the far north during the winter and in the eastern United States during the late summer.

When a Non-Fluid Acts Like a Fluid

Up to this point we have spoken primarily of convection in the atmosphere and the geosphere, but it is of importance also in the oceans. The miso soup example given earlier illustrates the movement of fluid, and hence of particles, that can occur when a convective cell is set up in a liquid.

Likewise, in the ocean convection—driven both by heat from the surface and, to a greater extent, by geothermal energy at the bottom—keeps the waters in constant circulation. Oceanic convection results in the transfer of heat throughout the depths and keeps the ocean stably stratified. In other words, the strata, or layers, corresponding to various temperature levels are kept stable and do not wildly fluctuate.

Ocean waters fit the most common, everyday definition of fluid, but as noted at the beginning of this essay, a fluid can be anything that flows—including a gas or, in special circumstances, a solid. Solid rocks or solid ice, in the form of glaciers, can be made to flow if the materials are deformed sufficiently. This occurs, for instance, when the weight of a glacier deforms ice at the bottom, thus causing the glacier as a whole to move. Likewise, geothermal energy can heat rock and cause it to flow, setting into motion the convective process of plate tectonics, described earlier, which literally moves the earth.

WHERE TO LEARN MORE

Educator's Guide to Convection (Web site). <http://www.solarviews.com/eng/edu/convect.htm>.

Erickson, Jon. *Plate Tectonics: Unraveling the Mysteries of the Earth.* New York: Facts on File, 1992.

Hess, Harry. "History of Ocean Basins" (Web site). <http://www-geology.ucdavis.edu/~GEL102/hess/jesse.htm>.

Jones, Helen. *Open-Ocean Deep Convection: A Field Guide* (Web site). <http://puddle.mit.edu/~helen/oodc.html>.

Ocean Oasis Teacher's Guide Activity 4 (Web site). <http://www.oceanoasis.org/teachersguide/activity4.html>.

Santrey, Laurence, and Lloyd Birmingham. *Heat.* Mahwah, NJ: Troll Associates, 1985.

Scorer, R. S., and Arjen Verkaik. *Spacious Skies.* Newton Abbot, England: David and Charles, 1989.

Sigurdsson, Haraldur. *Melting the Earth: The History of Ideas on Volcanic Eruptions.* New York: Oxford University Press, 1999.

Skinner, Brian J., Stephen C. Porter, and Daniel B. Botkin. *The Blue Planet: An Introduction to Earth System Science.* 5th ed. New York: John Wiley and Sons, 1999.

Smith, David G. *The Cambridge Encyclopedia of Earth Sciences.* New York: Cambridge University Press, 1981.

ENERGY AND EARTH

CONCEPT

Earth is a vast flow-through system for the input and output of energy. The overwhelming majority of the input to Earth's energy budget comes from the Sun in the form of solar radiation, with geothermal and tidal energy rounding out the picture. Each form of energy is converted into heat and re-radiated to space, but the radiation that leaves Earth travels in longer wavelengths than that which entered the planet. This is in accordance with the second law of thermodynamics, which shows that energy output will always be smaller than energy input and that energy which flows through a system will return to the environment in a degraded form. Yet what the Earth system does in processing that energy, particularly the portion that passes through the biosphere, is amazing. Some biological matter decays and, over the course of several hundred million years, produces fossil fuels that have given Earth a slight energy surplus. Human use of fossil fuels is rapidly depleting those sources, however, while posing new environmental problems, and this has encouraged the search for alternative forms of energy. Many of those forms, most notably geothermal energy, come from Earth itself.

HOW IT WORKS

Energy, Work, and Power

Physicists define energy as the ability of an object (and in some cases a nonobject, such as a magnetic force field) to accomplish work. "Work" in this context does not have the same meaning as it does in everyday life; along with the closely related concept of power, it is defined very specifically in a scientific context.

Work is the exertion of force over a given distance, and therefore it is measured in units of force multiplied by units of length. In the English system used by most Americans, a pound is the unit of force, and the foot-pound (ft-lb) would be the unit of work. However, scientists worldwide use SI, or the International System, which applies metric units. The metric unit of force is the newton (N), and the metric unit of work is the joule (J), equal to 1 newton-meter ($N \times m$).

Power is the rate at which work is accomplished over time and therefore is measured in units of work divided by units of time. The metric unit of power is the watt (W), named after James Watt (1736–1819), the Scottish inventor who developed the first fully viable steam engine and thus helped inaugurate the Industrial Revolution. A watt is equal to 1 J per second, but this is such a small unit that kilowatts, or units of 1,000 W, are more frequently used. Discussing the vast energy budget of Earth itself, however, requires use of an even larger unit: the terawatt (TW), equal to 10^{12} (one trillion) W.

Ironically, Watt himself—like most people in the British Isles and America—lived in a world that used the British system, in which the unit of power is the foot-pound per second. The latter unit, too, is very small, so for measuring the power of his steam engine, Watt suggested a unit based on something quite familiar to the people of his time: the power of a horse. One horsepower (hp) is equal to 550 ft-lb per second.

In the present context, we will rely as much as possible on SI units, especially because the watt is widely used in America. Horsepower typically is applied in the United States only for measuring the power of a mechanical device, such as an automobile or even a garbage disposal. For measuring electrical power, particularly in larger quantities, the SI kilowatt (kW) is used. When an electric utility performs a meter reading on a family's power usage, for instance, it measures that usage in terms of electrical "work" performed for the family and thus bills them by the kilowatt-hour (kWh).

Varieties of Energy

In the most fundamental sense, there are only three kinds of energy: kinetic, potential, and mass, or rest, energy. These types are, respectively, the energy an object possesses by virtue of its motion, its position (or its ability to perform work), and its mass. The first two are understood in relation to each other: for example, a ball held over the side of a building has a certain gravitational potential energy, but once it is dropped, it begins to lose potential energy and gain kinetic energy. The faster it moves, the greater the kinetic energy; but as it covers more distance, the less its potential energy is. (See Earth Systems for more about the kinetic-potential energy system.)

As Earth moves around the Sun, the gravitational interaction between the two bodies is not unlike that of the baseball and the ground in the illustration just given. Earth makes an elliptical, or oval-shaped, path in its orbit, meaning that the distance between it and the Sun is not uniform. At its furthest distance, Earth's potential energy is maximized, but as it comes closer to the Sun in its orbital path, its kinetic energy increases, with a corresponding decrease of potential energy.

MASS AND ENERGY. Mass, or rest, energy is identified in the famous formula $E = mc^2$, derived by Albert Einstein (1879–1955). In simple terms Einstein's formula means that every object possesses an amount of energy equal to its mass multiplied by the speed of light squared. Given that light travels at 186,000 mi. (299,339 km) per second, this is an enormous figure, even for a small object. A mere baseball, which weighs about 0.333 lb. (0.15 kg), possesses enough energy to yield about 3.75 billion kWh worth of power—enough to run all the lights and appliances in a typical American home for more than 156,000 years!

To release this energy in significant quantities, it would be necessary to accelerate the baseball to a speed close to that of light. Even in ordinary experience, however, very small amounts of mass are converted to energy. For instance, when a fire burns, the mass of the ashes combined with that of the particles and gases sent into the atmosphere is smaller (by an almost imperceptible fraction) than the mass of the original wood. The "lost" mass is converted to energy. These mass-energy conversions occur on a much larger level in nuclear reactions, such as the nuclear fusion of hydrogen atoms to form helium in the solar core (see Sun, Moon, and Earth).

MANIFESTATIONS OF ENERGY. In discussing kinetic and potential energy, the example of dropping a baseball from a height illustrates these two types of energy in a gravitational field—that is, the gravitational field of Earth. Yet the concept of potential and kinetic energy translates to a situation involving an electromagnetic field as well. For instance, the positive or negative attraction between two electromagnetically charged particles is analogous to the force of gravity, and a system of two or more charges possesses a certain amount of kinetic and potential electromagnetic energy.

Electromagnetic energy, which is the form in which solar power reaches Earth, is a type of energy that (as its name suggests) combines both electrical and magnetic energy. Another important form of energy in the Earth system is thermal, or heat, energy, which is the kinetic energy of molecules, since heat is simply the result of molecular motion.

Other types of energy include sound, chemical, and nuclear energy. Sound waves, which require a physical medium such as air in which to travel, are simply pressure fluctuations that carry varying levels of energy, depending on the frequency (pitch) and amplitude (volume) of the waves. Chemical energy makes possible the forming and releasing of molecular bonds, and, for this reason, chemical reactions often are accompanied by the production of heat. Whereas chemical energy concerns the bonds *between* atoms, nuclear energy relates to the bonds *within* them. Nuclear fission reactions involve the splitting of an atomic nucleus, while nuclear fusion is the joining of nuclei.

Heat and Thermodynamics

Thermodynamics is the study of the relationships between heat, work, and energy. As with work, energy, and power, heat and the related concept of temperature are terms that have special definitions in the physical sciences. Heat itself is not to be confused with thermal energy, which, as noted earlier, is the kinetic energy that arises from the motion of particles at the atomic or molecular level. The greater the movement of these particles relative to one another, the greater the thermal energy.

Heat is internal thermal energy that flows from one body of matter to another. It is not the same as the energy contained in a system—that is, the internal thermal energy of the system. Rather than being "energy-in-residence," heat is "energy-in-transit." This may seem a little confusing, but all it means is that heat, in its scientific sense, exists only when internal energy is being transferred. As for temperature, it is not (as is commonly believed) a measure of heat and cold. Instead, temperature indicates the direction of internal energy flow between bodies and the average molecular kinetic energy in transit between those bodies.

In any case, temperature could not be a measure of heat and cold, as though these two were equal and opposing entities, because, scientifically speaking, there is no such thing as cold—only an absence of heat. When we place an ice cube in a cup of coffee, we say that the ice is there to "cool the coffee down," but, in fact, the opposite is happening: the coffee is warming up the ice cube, and in the process of doing so, it loses heat. This may seem like a difference of semantics, but it is not. It is a physical law that the flow of heat is always from a high-temperature reservoir to a low-temperature reservoir. Even air conditioners and refrigerators work by pulling heat out of a compartment rather than by bringing cold *in.*

MEASURING TEMPERATURE AND HEAT. Temperature, of course, can be measured by either the Fahrenheit or the Centigrade scales familiar in everyday life. Scientists, however, prefer the Kelvin (K) scale, established by William Thomson, Lord Kelvin (1824–1907). Drawing on the discovery made by the French physicist and chemist J. A. C. Charles (1746–1823) that gas at 0°C (32°F) regularly contracts by about 1/273 of its volume for every Celsius degree drop in temperature, Thomson derived the value of absolute zero (the temperature at which molecular motion virtually ceases) as –273.15°C (–459.67°F). The Kelvin and Celsius scales are thus directly related: Celsius temperatures can be converted to Kelvin units (for which neither the word nor the symbol for "degree" is used) by adding 273.15.

Heat, on the other hand, is measured not by degrees but by the same units as work. Energy is the ability to perform work, so heat or work units are also units of energy. Aside from the joule, heat often is measured by the kilocalorie, or the amount of heat that must be added to or removed from 1 kg of water to change its temperature by 1°C. As its name suggests, a kilocalorie is 1,000 calories, a calorie being the amount of heat required to change the temperature in 1 g of water by 1°C. The dietary calorie with which most people are familiar, however, is the same as a kilocalorie.

THE LAWS OF THERMODYNAMICS

The three laws of thermodynamics collectively show that it is impossible for a system to produce more energy than was put into it or even to produce an equal amount of usable energy. In other words, a perfectly efficient system—whether an engine or the entire Earth—is an impossibility. Derived during a period of about 60 years beginning in the 1840s, the laws of thermodynamics helped scientists and engineers improve the machines that powered the height of the Industrial Age. They also revealed the impossibility of constructing anything approaching a perpetual-motion machine, that great quest of dreamers over the ages, which the laws of thermodynamics proved to be an impossible dream.

THE FIRST LAW OF THERMODYNAMICS. The first law of thermodynamics is related to the conservation of energy, a physical law whereby the total energy in a system remains the same, though transformations of energy from one form to another take place. Such transformations occur frequently in the Earth system, as when a plant receives electromagnetic energy from the Sun and converts it to chemical potential energy in the form of carbohydrates. Likewise humans, by building dams, can harness the gravitational potential energy of flowing water and convert it into electromagnetic energy.

The conservation of energy, in effect, states that "the glass is half full," meaning that we can obtain as much energy from a system as we put into it. While saying the same thing, the first law of thermodynamics in effect states that "the glass is half empty,": that is, that we can obtain *no more* energy from a system than we put into it. According to this law, because the amount of energy in a system remains constant, it is impossible to perform work that results in an energy output greater than the energy input.

The term *law* in the physical sciences is no empty expression; it means that a principle has been shown to be the case always and may be expected to remain the case in all situations. It is possible, of course, for a physical law to be overturned in light of later evidence. It is not likely, however, that any set of circumstances in the universe will ever disprove the core truth behind this law, which may be stated colloquially as "You can't get something for nothing."

THE SECOND LAW OF THERMODYNAMICS. In a 1959 lecture published as *The Two Cultures and the Scientific Revolution,* the British writer and scientist C. P. Snow (1905–1980) compared transfers of heat and energy to a game. The laws of thermodynamics are its rules, and, as Snow stated, the first law proves that it is impossible to win at this game, while the second law shows the impossibility of breaking even.

The second law of thermodynamics is more complicated than the first and is stated in a number of ways, though they are all interrelated. According to this law, spontaneous or unaided transfers of energy are irreversible and impossible without an increase of entropy in the universe. Entropy is the tendency of natural systems toward breakdown, specifically, the tendency for the energy in a system to be dissipated or degraded. (Later in this essay, we discuss examples of energy that has been degraded—for instance, wood that has been burned to produce ashes.) The second law means that spontaneous processes are irreversible and that it is impossible, without the additional input of energy, to transfer heat from a colder to a hotter body or to convert heat into an equal amount of work.

Whereas the first law showed engineers the impossibility of building a perpetual-motion machine, the second law proves that it is impossible to build even a perfectly efficient engine. Of all the energy we put into our automobiles in the form of gasoline (which is chemical potential energy in the form of hydrocarbons derived from the fossilized remains of dinosaurs in the earth), only about 30% of it goes into moving the car forward. The rest is dissipated in a number of ways, chiefly through heat and sound. Entropy, as it turns out, is inescapable and as inevitable as death. In fact, death itself is a result of entropy in the systems of all living things.

THE THIRD LAW OF THERMODYNAMICS. The third law of thermodynamics is not as well known as the other two and has little bearing on the discussion at hand, but it deserves at least brief mention. According to the third law, at the temperature of absolute zero entropy also approaches zero, which might sound like a way out of the restrictions imposed by the first two laws. All it really means is that absolute zero is impossible to reach—or, as Snow put it, the third law shows that "you can't escape the game."

In 1824 the French physicist and engineer Nicolas Léonard Sadi Carnot (1796–1832) had shown that an engine could achieve maximum efficiency if its lowest operating temperature were absolute zero. His work influenced that of Kelvin, who established the absolute-temperature scale mentioned earlier. Additionally, Carnot's discoveries informed the development of the third law. Whereas the second law is not derived from the first (though it is certainly consistent with it), the third law relies on the second: if it is impossible to build a perfectly efficient engine, as the second law states, it is likewise impossible to reach absolute zero.

This, of course, has not stopped scientists from attempting to achieve absolute zero, most properly defined as the temperature at which the motion of the average atom or molecule is zero. Helium atoms, in fact, never fully cease their motion, even at temperatures very close to 0K—and scientists have come very, very close. In 1993 physicists at the Helsinki University of Technology Low Temperature Laboratory in Finland used a nuclear demagnetization device to achieve a temperature of 2.8×10^{-10}K, or 0.00000000028K. This amounts to a difference of only 28 parts in 100 billion between that temperature and absolute zero.

EARTH'S ENERGY INPUT

Just as households have financial budgets, a system such as Earth (see Earth Systems) has an energy budget. The latter may be defined as the total amount of energy available to a system or, more specifically, the difference between the

THE POHUTU GEYSER AT THE ROTORUA WHAKAREWAREWA THERMAL AREA IN NORTH ISLAND, NEW ZEALAND. *(© A. N. T./Photo Researchers. Reproduced by permission.)*

energy flowing into the system and the energy lost by it. Having reviewed the laws of thermodynamics, one might suspect that a great deal of energy is dissipated in the operation of the Earth system, and, of course, that is absolutely correct.

Earth receives 174,000 TW of energy, or 174 quadrillion J per second. Human civilization, by contrast, uses only 10 TW, or about 0.00574% as much as Earth's total energy. Of that total, there are three principal sources, though one of these sources—the Sun—dwarfs the other two in importance. The breakdown of Earth's energy input, along with the percentage of the total that each portion constitutes, is as follows:

- Solar radiation: 99.985%
- Geothermal energy: 0.013%
- Tidal energy: 0.002%

SOLAR RADIATION. The Sun radiates electromagnetic energy, which, as mentioned previously, is a form of energy that produces both electric and magnetic fields. Electromagnetic energy travels in waves, and since waves follow regular patterns, it is possible to know that those waves with shorter wavelengths have a higher frequency and thus higher energy levels.

The electromagnetic spectrum contains a variety of waves, each with progressively higher energy levels, including long-wave and short-wave radio; microwaves (used for TV transmissions); infrared, visible, and ultraviolet light; x rays; and ultra-high-energy gamma rays. Visible light is only a very small portion of this spectrum, and each color has its own narrow wavelength range.

Red has the least energy and purple or violet the most; hence, the names *infrared* for light with less energy than red, and *ultraviolet* for light with more energy than violet. The order of these wavelengths of light, along with the colors between, is remembered easily by the mnemonic device ROY G. BIV (standing for red, orange, yellow, green, blue, indigo, and violet). (Actually, there are only six major color ranges, and the name really should be *ROY G. BV.*)

Although it covers the entire electromagnetic spectrum, energy from the Sun is referred to by earth scientists as short-wavelength radiation. This is because the solar energy that enters the Earth system is shorter in wavelength (and thus higher in energy level) than the energy returned to space by Earth. (We discuss the degradation of energy in the Earth system later in this essay.) Without solar radiation, the life-giving processes of the hydrosphere, biosphere, and atmosphere would be impossible. An example is photosynthesis, the biological conversion of electromagnetic energy to chemical energy in plants. (See the later discussion of photosynthesis and the food web.)

GEOTHERMAL ENERGY. A much smaller, but still significant component of Earth's energy budget is geothermal energy, the planet's internal heat energy. Much of this heat comes from Earth's core, which has temperatures as high as 8,132°F (4,500°C) and from whence thermal energy circulates throughout the planet's interior. Also significant is the heat from radioactive elements, most notably uranium and thorium, near Earth's surface.

This thermal energy heats groundwater, and thus the principal visible sources of geothermal energy include geysers, hot springs, and fumaroles—fissures, created by volcanoes, from which hot gases pour. There are several types of geothermal energy reserves, among them dry and wet steam fields. The first of these reserves occurs when groundwater boils normally, whereas in the second type of reserve, groundwater is superheated, or prevented from boiling even though its temperature is above the boiling point. In both cases the waters have a much higher concentration of gases and minerals than ordinary groundwater. Another type of reserve can be found under the ocean floors, where natural gas mixes with very hot water.

Geothermal energy powers seismic activity as well as volcanic eruptions and mountain building, which together have played a significant role in shaping Earth as we know it today. Aside from its obvious impact on the planet's terrain, geothermal energy has had an indirect influence on the transfer of vital elements from beneath Earth's surface, a benefit of volcanic activity. (See the later discussion of the human use of geothermal energy in this essay.)

TIDAL ENERGY. Whereas the principal form of energy in Earth's budget comes from the Sun and the secondary source from Earth itself, the third type of energy input to the Earth system comes chiefly from the Moon. The Sun also affects tides, but because of its close proximity to Earth, the Moon has more influence over the movements of our planet's ocean waters.

Though the Moon is much smaller than Earth, it is larger, in proportion to the planet it orbits, than any satellite in the solar system (with the possible exception of Pluto's moon Charon). Given this fact, combined with its close proximity to Earth, it is understandable that the Moon would exert a powerful pull on its host planet. The gravitational pull of the Moon (and, to a lesser extent, that of the Sun) on Earth causes the oceans to bulge outward on the side of Earth closest to the Moon. At the same time, the oceans on the opposite side of the planet bulge in response. (See Sun, Moon, and Earth for more about tides and the bulges that result from the Moon's gravitational pull.)

This gravitational pull creates a torque that acts as a brake on Earth's rotation, producing a relatively small amount of energy that is dissipated primarily within the waters of the ocean. Incidentally, the lunar-solar tidal torque, by increasing the amount of time it takes Earth to turn on its axis, is causing a gradual increase in the length of a day. Today, of course, there are 365.25 days in a year, but about 650 million years ago there were 400 days. In other words, Earth made 400 revolutions on its axis in the period of time it took it to revolve around the Sun. The change is a result of the fact that Earth's rotation is being slowed by 24 microseconds a year.

Energy Profit and Loss

Focusing now on solar radiation, since it is by far the greatest source of energy input to the Earth system, let us consider Earth's energy budget in terms of "profit and loss." In other words, how much useful energy output is denied to Earth owing to the laws of thermodynamics and other factors?

First of all, a good 30% of the Sun's energy input is reflected back into space unchanged, without entering Earth's atmosphere. This results from our planet's *albedo,* or reflective power. Albedo is the proportion of incoming radiation that is reflected by a body (e.g., a planet) or surface such as a cloud: the higher the proportion of incoming radiation that a planet deflects, the higher its albedo. The latter is influenced by such factors as solar angle, amount of cloud cover, particles in the atmosphere, and the character of the planetary surface.

Another 25% of solar radiation is absorbed by the atmosphere, while about 45% is absorbed at the planetary surface by living and nonliving materials. Thus, electromagnetic energy from the Sun enters the atmosphere, biosphere, and hydrosphere, where it is converted to other forms of energy, primarily thermal. Some of this thermal energy, for instance, causes the evaporation of water, which cycles through the atmosphere and then reenters the hydrosphere as precipitation. In other cases, absorbed radiation drives atmospheric and hydrologic distribution mechanisms, including winds, water currents, and waves. A very small, but extremely significant portion of incoming solar radiation goes into plant photosynthesis, discussed later.

ENERGY DEGRADATION. The energy that enters the Earth system—not only solar radiation but also geothermal and tidal energy—ultimately leaves the system. As shown by the second law of thermodynamics, however, the energy that departs the Earth system will be in a degraded form compared with the energy that entered it.

In a steam engine, water in the form of steam goes to work to power gears or levers. In the process, it cools, and the resulting cool water constitutes a degraded form of energy. Likewise, the ashes that remain after a fire or the fumes that are a by-product of an internal combustion engine's operation contain degraded forms of energy compared with that in the original wood or gasoline, respectively. In the same way, Earth receives short-wavelength energy from the Sun, but the energy it radiates to space is in a long-wavelength form.

All physical bodies with a temperature greater than absolute zero emit electromagnetic energy in accordance with their surface temperatures, and the hotter the body, the shorter the wavelength of the radiation. The sunlight that enters Earth's atmosphere is divided between the visible portion of the spectrum and the high-frequency side of the infrared portion. (Note that the Sun emits energy across the entire electromagnetic spectrum, but only a small part gets through Earth's atmospheric covering.) Earth, with an average surface temperature of 59°F (15°C), is much cooler than the Sun, with its average surface temperature of about 10,000°F (5,538°C). The radiation Earth sends back into space, then, is on the low-frequency, long-wavelength side of the infrared spectrum.

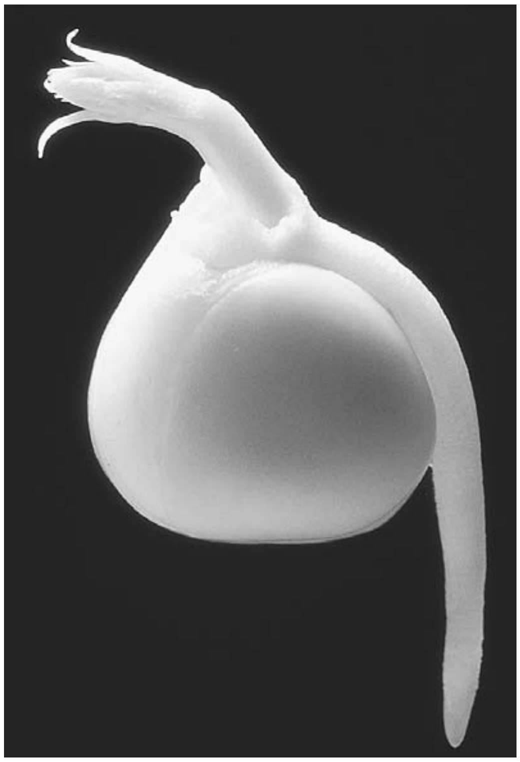
NO ENERGY LOSS. In accordance with the conservation of energy, no energy truly has been lost. In a relatively simple system, such as an automobile, chemical potential energy enters the vehicle in the form of gasoline and, after being processed by the engine, exits in a variety of forms. There is the kinetic energy that turns the wheels; the thermal energy of the engine and exhaust; electromagnetic energy from the battery for the headlights, dashboard lights, radio, air conditioning, and so on; and the sound energy dissipated in the noise of the car. If one could add all those energy components together, one would find that all the energy that entered the system left the system. Note, however, that once again the process is irreversible: one can use gasoline to power a battery and hence a car radio, but the radio or the battery cannot generate gasoline.

The Earth system is much more complex, of course, but the same principle applies: about 174,000 TW of energy enter the system, and about 174,000 TW are used in the form of heat. Of the portion that enters the atmosphere, some goes into warming the planet, some into moving the air and water, and a very small part into the all-important biological processes described later, but *all* of it is used. It should be noted that Earth has a very small energy surplus, owing to the accumulation of undecomposed biomass that ultimately becomes fossil fuels; however, the amount of energy involved is minor compared with the larger energy budget. (For more information about biomass and fossil fuels, see the following discussion.)

THE GREENHOUSE EFFECT. Not only is there no net loss of energy in the universe, but Earth itself also possesses a remarkably efficient system for making use of the energy it receives. This is the greenhouse effect, whereby the planet essentially recycles the degraded energy it is in the process or returning to space.

Water vapor and carbon dioxide, as well as methane, nitrous oxide, and ozone, all absorb long-wavelength radiated energy as the latter makes its way up through the atmosphere. When heated, these radiatively active gases (as they are called) re-radiate the energy, now at even longer wavelengths. In so doing, they slow the planet's rate of cooling. Without the greenhouse effect, surface temperatures would be about 50°F (10°C) cooler than they are—that is, around 17.6°F (–8°C). This, of course, is well below the freezing temperature of water and much too cold for Earth's biological processes. Thus, the greenhouse effect literally preserves life on the planet.

It may be surprising to learn that the greenhouse effect, a term often heard in the context of dire environmental warnings, is a natural and healthful part of the Earth system. Like many useful things, the greenhouse effect is not necessarily better in larger doses, however, and that is the problem. It is believed that human activities have resulted in an increase of radiatively active gases, which could lead to global warming. Among such gases are the chlorofluorocarbons (CFCs) used in aerosol cans, now banned by international treaty. Also of concern are the high levels of carbon dioxide produced by the burning of fossil fuels.



A GERMINATING GARDEN PEA SEEKS SUNLIGHT FOR PHOTOSYNTHESIS. *(© J. Burgess/Photo Researchers. Reproduced by permission.)*

REAL-LIFE APPLICATIONS

ENERGY AND THE LIVING WORLD

One of the most interesting components of the entire energy picture is the relationship of energy input, energy output, and the biosphere. Particularly important is the use of solar radiation for the purposes of photosynthesis, an activity that constitutes a small but vital sector of Earth's energy budget.

Though it accounts for only about 1% of the energy received from solar radiation, photosynthesis is essential to the sustenance of life. In photosynthesis, plants receive solar radiation, carbon dioxide, and water, which chemically react to produce carbohydrates and oxygen. Animals depend not only on the oxygen but also on carbohydrates, such as sugars, and thus photosynthesis makes possible the development of the food chains that constitute the world of living creatures.

FOOD WEBS. Even though *food chain* is a well-known expression, modern scientists prefer the term *food web,* because it more accurately portrays the complicated relationships involved. The term food chain, as it is commonly used, implies a strict hierarchical structure in which (to use another popular phrase) "the big fish eat the little fish." In fact, the relationship between participants is not quite so neatly defined.

It is, however, possible to describe a food web in terms of a few key players, or types of players. There are primary producers, which are green plants, and primary consumers—herbivores, or plant-eating animals. Secondary consumers (and those at further levels, such as tertiary and quaternary) are either carnivores—that is, meat-eating animals who eat the herbivores—or omnivores, which are both plant and animal eaters. For example, omnivorous humans eat herbivorous cows, who have eaten plants.

Carnivores and omnivores, however, are not really at the top of the food chain, so to speak; rather, in line with the non-hierarchical idea of the food web, they represent points in an interlocking set of relationships. Materials from plants, herbivores, carnivores, and omnivores ultimately will all be consumed by the lowliest of creatures, that is, detritivores, or decomposers, including bacteria and worms. At each stage energy is transferred, and, as always, the second law of thermodynamics comes into play. The energy is degraded in transfer; specifically, the further away an organism is from the original plant source, the less a given quantity of fuel contributes to its growth. It is interesting to note the economy of energy use at the detritivore stage.

Worms and other decomposers are exceedingly efficient feeders, working the same food particles over and over and extracting more stored energy each time. They then produce waste products that increase the vitamin content of the soil, thus enabling the growth of plants and the continuation of the biological cycle. Also, detritivores may contribute directly (and, from the larger energy-cycle perspective, less efficiently) to providing fuel for carnivores, as when a bird eats a worm.

BIOENERGY

The food web is the mechanism whereby energy is cycled through the biosphere, as fuel in the form of food. Plant matter and other biological forms also can serve as direct sources of heat, providing fuel that can be either burned without processing or converted into gas or alcohol. In such situations the plant matter is described as bioenergy, or energy derived from biological sources that are used as fuel.

Materials that are burned or processed to produce bioenergy are called biomass. Examples of the latter include wood logs burned on a fire, probably the oldest type of fuel known to humankind and still one of the principal forms of heating available in many developing countries. In fact, some of the least technologically advanced and most technologically advanced nations are alike in their use of another variety of biomass: waste.

Dried animal dung provides heating material in many a third world village where electricity is unknown, while at the other end of the technological spectrum, some Western municipalities extract burnable biomass from processed sewage. Since Western countries have at their disposal plenty of other energy sources, they typically burn off the methane gas and dried waste material from treated sewage simply as a means of removing it, rather than as an energy source. Those materials could provide usable energy, however.

The products of sewage treatment, of course, are the result of processing, which involves the conversion of biomass to either gases (for example, methane, as mentioned previously) or alcohol. Farmers in rural China, for instance, often place agricultural waste and sewage in small closed pits, from whence they extract burnable methane gas. In the United States, Brazil, and other countries with abundant farmland, some of the agricultural output is directed not toward production of food but toward production of fuel in the form of ethanol, a type of alcohol made from sugarcane, corn, or sorghum grain. Ethanol can be mixed with gasoline to run an automobile or burned alone in specially modified engines. In either case, the fuel burns much more cleanly, producing less poisonous carbon monoxide than ordinary gasoline.

FOSSIL FUELS

Biomass is potentially renewable, whereas another source of bioenergy is not. This is the bioenergy from fossil fuels—buried deposits of petroleum, coal, peat, natural gas, and other organic compounds. (Actually, fossil fuels typically are considered separate from other forms of bioenergy, because they are nonrenewable.)

Fossil fuels are the product of plants and animals that lived millions of years ago, died, decomposed, and became part of Earth's interior. Over the ages, as more sediment weighed down on these organic deposits, the weight applied more pressure, generated more heat, and led to the concentration of this decomposed material, which became a valuable source of energy.

USING UP RESOURCES. Given the vast spans of time that have passed, as well as the almost inconceivable numbers of plants and animals, the supplies of fossil-fuel energy stored under Earth's surface are enormous. But they are *not* infinite, and, as noted earlier, they are nonrenewable; once they are gone, they might as well be gone forever, because it would take hundreds of millions of years to produce additional deposits. Furthermore, humans are using up these energy sources at an alarming rate.

Up until about 1750, Earth's fossil-fuel deposits were largely intact, but after that time industrialized societies began to extract coal for heating, transportation, and industrial uses. Today it remains a leading means of generating electrical power. During the twentieth century, petroleum increasingly was directed toward transportation, and this led to the extraction of still more of Earth's fossil-fuel deposits. If civilization continues to consume these products at its current rate, reserves will be exhausted long before the end of the twenty-first century.

SEARCHING FOR ENERGY SOURCES. There are several reasons not to panic over the loss of fossil-fuel resources. First, known reserves are just that—they are the ones that energy companies, and their geologists, know about at the present time. As long as plentiful resources are available, corporations and governments do not feel a pressing need to search for more, but as those resources are used up, such searches become economically necessary. For a time at least, these searches will continue to yield new (though increasingly harder to reach) deposits.

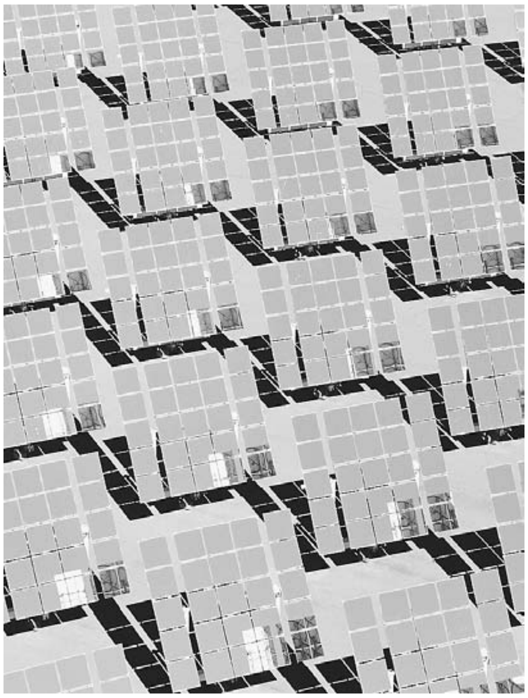
Also, fears about Earth running out of fuel are built on the assumption that no one will develop other sources of energy, a few of which are discussed at the conclusion of this essay. Though most of the known alternative energy sources face their own challenges, it is thoroughly conceivable that scientists of the future will develop means of completely replacing fossil fuels as the source of electrical power, fuel for transport, and other uses. After all, there was once a time (during the second half of the nineteenth century) when Americans became increasingly anxious over dwindling reserves of a vital energy resource essential for powering the nation's lamps: whale oil.

ENVIRONMENTAL CONCERNS

While the loss of energy reserves may not be an immediate cause for panic, the effects of fossil-fuel burning on the environment have alarmed scientists and environmentalists alike. At the most basic level, there is the environmental impact posed by the extraction of fossil fuels. Coal mines, for instance, have been used up and now sit abandoned, their land worthless for any purpose. There is also the environmental danger created by hazards in misuse or transport of fossil fuels—for example, the vast oil spill caused by the grounding of the Exxon *Valdez* near Alaska's Prince William Sound in 1989.

By far the greatest environmental concern raised by fossil fuels, however, is the effect they produce in the atmosphere when burned. For instance, one of the impurities in coal is sulfur, and when coal burns, the sulfur reacts with oxygen in the combustion process to create sulfur dioxide and sulfur trioxide. Sulfur trioxide reacts with water in the air, creating sulfuric acid and thus acid rain, which can endanger plant and animal life as well as corrode metals and building materials.

THE GREENHOUSE EFFECT REVISITED. Even greater fears center on the release into the atmosphere of carbon monoxide and carbon dioxide, both of which are by-products in the burning of fossil fuels. The first is a poison, whereas the second is a vital part of the life cycle, yet carbon dioxide, in fact, may pose the greater threat.



AN ARRAY OF SOLAR REFLECTORS NEAR ALBUQUERQUE, NEW MEXICO, HARNESSES SOLAR POWER. *(© J. Mead/Photo Researchers. Reproduced by permission.)*

Earlier we discussed the greenhouse effect, which, when it occurs naturally, is important to the preservation of life on the planet. The large number of internal-combustion engines in operation on Earth today produce an inordinate amount of carbon dioxide, which in turn provides the atmosphere with more radiatively active gas than it needs. According to many environmentalists, the result is, or will be, global warming. If it takes place over a long period of time, global warming could bring about serious hazards—in particular, the melting of the polar ice caps. It should be noted, however, that not all scientists are in agreement that global warming is occurring or that humans are the principal culprits inducing these environmental changes.

ALTERNATIVE ENERGY SOURCES

Whatever the merits of the various sides in the debate on global warming or the exhaustion of fuel resources, one need hardly be an environmentalist to agree that the world cannot forever rely only on existing fossil fuels and the technology that uses them. Today, even as scientists in laboratories around the world work to develop viable alternative means of powering industry, utilities, and transportation, many alternative energy sources are already in use.

Some of these energy forms are very old, for instance, burnable biomass, water power, or wind power, all of which date back to ancient times. Others are extremely new in concept, most notably, nuclear energy, which was developed in the twentieth century. Still others are new, high-tech versions of old-fashioned energy sources, the best example being solar power. In fact, all three major contributors to Earth's energy budget—solar radiation, tidal energy, and geothermal energy (discussed further later in this essay)—have been harnessed by human societies.

HIGH-TECH ENERGY SOLUTIONS. Several of the more ambitious ideas for energy creation, while they may capture people's imaginations, have significant drawbacks. There is the proposal, for instance, to extract hydrogen gas from water by means of electrolysis, potentially providing an extremely clean, virtually limitless source of energy. Hydrogen gas, however, is highly flammable, as the 1937 explosion of the airship *Hindenburg* illustrated, and, in any case, the fuel to provide the electricity necessary for electrolysis would have to come from somewhere, presumably, the burning of fossil fuels.

Nuclear energy, of course, has frightened many people in the wake of such well-known disasters as those that occurred at Three Mile Island, Pennsylvania, in 1979 and Chernobyl, in the former Soviet Union, in 1986. In fact, those two situations illustrate more about governments than they do about technology itself. No one died at Three Mile Island, and with an open society and media access to the site, the public outcry became so great that the plant was closed. By contrast, the Soviets' outmoded technology helped bring about the Chernobyl disaster, and the communist dictatorship's practice of censorship and suppression led to a massive cover-up that greatly increased the death toll. As a result, thousands died at Chernobyl, and thousands more died as the result of the indirect effects of nuclear pollution in the environment.

There is no question, however, that nuclear energy does pose an enormous potential environmental threat from its waste products. Spent fuel rods, if simply buried, eventually leak radioactive waste into the water table and could kill or harm vast populations. This all relates to nuclear *fission,* the only type of peaceful nuclear

KEY TERMS

ABSOLUTE ZERO: The temperature at which all molecular motion virtually ceases.

ALBEDO: The reflective power of a surface or body or, more specifically, the proportion of incoming radiation that the surface or body reflects.

ATMOSPHERE: In general, an atmosphere is a blanket of gases surrounding a planet. Unless otherwise identified, however, the term refers to the atmosphere of Earth, which consists of nitrogen (78%), oxygen (21%), argon (0.93%), and other substances that include water vapor, carbon dioxide, ozone, and noble gases such as neon, which together comprise 0.07%.

BIOENERGY: Energy derived from biological sources that are used directly as fuel (as opposed to food, which becomes fuel).

BIOMASS: Materials that are burned or processed to produce bioenergy.

BIOSPHERE: A combination of all living things on Earth—plants, mammals, birds, reptiles, amphibians, aquatic life, insects, viruses, single-cell organisms, and so on—as well as all formerly living things that have not yet decomposed.

CALORIE: A measure of heat or energy in the SI, or metric, system, equal to the heat that must be added to or removed from 1 g of water to change its temperature by 1°C. The dietary calorie with which most people are familiar is the same as the kilocalorie, or 1,000 calories.

CONSERVATION OF ENERGY: A law of physics that holds that within a system isolated from all outside factors, the total amount of energy remains the same, though transformations of energy from one form to another take place. The first law of thermodynamics is the same as the conservation of energy.

ELECTROMAGNETIC ENERGY: A form of energy with electric and magnetic components, which travels in waves.

ELECTROMAGNETIC SPECTRUM: The complete range of electromagnetic waves on a continuous distribution from a very low range of frequencies and energy levels, with a correspondingly long wavelength, to a very high range of frequencies and energy levels, with a correspondingly short wavelength. Included on the electromagnetic spectrum are long-wave and short-wave radio; microwaves; infrared, visible, and ultraviolet light; x rays; and gamma rays.

ENERGY: The ability of an object (or in some cases a nonobject, such as a magnetic force field) to accomplish work.

ENERGY BUDGET: The total amount of energy available to a system or, more specifically, the difference between the energy flowing into the system and the energy lost by it.

ENTROPY: The tendency of natural systems toward breakdown and, specifically, the tendency for the energy in a system to be dissipated. Entropy is related closely to the second law of thermodynamics.

ENVIRONMENT: In discussing systems, the term environment refers to the surroundings—everything external to and separate from the system.

KEY TERMS CONTINUED

FIRST LAW OF THERMODYNAMICS: A law of physics stating that the amount of energy in a system remains constant, and therefore it is impossible to perform work that results in an energy output greater than the energy input. This is the same as the conservation of energy.

FOSSIL FUELS: Nonrenewable forms of bioenergy, including petroleum, coal, peat, natural gas, and other organic compounds usable as fuel.

FREQUENCY: The number of waves, measured in Hertz, passing through a given point during the interval of one second. The higher the frequency, the shorter the wavelength.

GEOTHERMAL ENERGY: Heat, or thermal, energy from Earth's interior.

GREENHOUSE EFFECT: Warming of the lower atmosphere and surface of Earth. This occurs because of the absorption of long-wavelength radiation from the planet's surface by certain radiatively active gases, such as carbon dioxide and water vapor, in the atmosphere. These gases are heated and ultimately re-radiate energy at an even longer wavelength to space.

HEAT: Internal thermal energy that flows from one body of matter to another.

HERTZ: A unit for measuring frequency, equal to one cycle per second. High frequencies are expressed in terms of kilohertz (kHz; 10^3 or 1,000 cycles per second), megahertz (MHz; 10^6 or one million cycles per second), and gigahertz (GHz; 10^9 or one billion cycles per second.)

HYDROSPHERE: The entirety of Earth's water, excluding water vapor in the atmosphere but including all oceans, lakes, streams, groundwater, snow, and ice.

JOULE: The SI measure of work. One joule (J) is equal to the work required to accelerate 1 kg of mass by 1 m per second squared (1 m/s^2) over a distance of 1 m. Owing to the small size of the joule, however, it often is replaced by the kilowatt-hour, equal to 3.6 million (3.6×10^6) J.

KELVIN SCALE: Established by William Thompson, Lord Kelvin (1824–1907), the Kelvin scale measures temperature in relation to absolute zero, or 0K. (Note that units in the Kelvin system, known as kelvins, do not include the word or symbol for "degree.") The Kelvin scale, which is the system usually favored by scientists, is related directly to the Celsius scale; hence, Celsius temperatures can be converted to kelvins by adding 273.15.

KINETIC ENERGY: The energy that an object possesses by virtue of its motion.

LAW: A scientific principle that is shown always to be the case and for which no exceptions are deemed possible.

MASS ENERGY: The energy an object possesses by virtue of its mass. Sometimes called *rest energy.*

NUCLEAR FISSION: A nuclear reaction that involves the splitting of an atomic nucleus.

NUCLEAR FUSION: A nuclear reaction that involves the joining of atomic nuclei.

NUCLEUS: The center of an atom, a region where protons and neutrons are located and around which electrons spin.

KEY TERMS CONTINUED

PHOTOSYNTHESIS: The biological conversion of light energy (that is, electromagnetic energy) to chemical energy in plants.

POTENTIAL ENERGY: The energy that an object possesses by virtue of its position or its ability to perform work.

POWER: The rate at which work is accomplished over time, a figure rendered mathematically as work divided by time. The SI unit of power is the watt, while the British unit is the foot-pound per second.

RADIATION: The transfer of energy by means of electromagnetic waves, which require no physical medium (for example, water or air) for the transfer. Earth receives the Sun's energy via the electromagnetic spectrum by means of radiation.

RADIOACTIVITY: A term describing a phenomenon whereby certain materials are subject to a form of decay brought about by the emission of high-energy particles or radiation. Forms of particles or energy include alpha particles (positively charged helium nuclei), beta particles (either electrons or subatomic particles called positrons), or gamma rays, which occupy the highest energy level in the electromagnetic spectrum.

SECOND LAW OF THERMODYNAMICS: A law of physics stating that spontaneous or unaided transfers of energy are irreversible and impossible without an increase of entropy in the universe. It is therefore impossible, without the additional input of energy, to transfer heat from a colder to a hotter body or to convert heat into an equal amount of work.

SI: An abbreviation of the French term *Système International d'Unités,* or International System of Units. Based on the metric system, SI is the system of measurement units in use by scientists worldwide.

SYSTEM: Any set of interactions that can be set apart mentally from the rest of the universe for the purposes of study, observation, and measurement.

TEMPERATURE: The direction of internal energy flow between two systems when heat is being transferred. Temperature measures the average molecular kinetic energy in transit between those systems.

TERAWATT: See *watt.*

THERMAL ENERGY: Heat energy, a form of kinetic energy produced by the motion of atomic or molecular particles in relation to one another. The greater the relative motion of these particles, the greater the thermal energy.

WATT: The metric unit of power, equal to 1 J per second. Because this is such a small unit, scientists and engineers typically speak in terms of kilowatts, or units of 1,000 W. Very large figures, such as those relating to Earth's energy budget, usually are given in terawatts, or 10^{12} (one trillion) W.

WAVELENGTH: The distance between a crest and the adjacent crest or the trough and an adjacent trough of a wave. Wavelength is related inversely to frequency, meaning that the shorter the wavelength, the higher the frequency.

WORK: The exertion of force over a given distance. In the metric, or SI, system, work is measured by the joule (J) and in the British system by the foot-pound (ft-lb).

power in use today. In building the hydrogen bomb in the mid-twentieth century, physicists and chemists used the much greater power of nuclear *fusion,* or the bonding of atomic nuclei. If nuclear fusion could be produced in controlled reactions, for peaceful use, it would provide safe, cheap, limitless power to the planet.

GEOTHERMAL ENERGY. Unless or until nuclear fusion or some more advanced alternative energy source is developed, societies will continue to rely on fossil fuels for the bulk of their energy needs. At the same time, alternative sources will continue to supply energy in certain situations. An excellent example of such a source is geothermal energy, harnessed by peoples in areas as widespread as New Zealand and Iceland centuries ago.

Long before the first Europeans arrived in New Zealand, the native Maori people used geothermal energy from geysers to cook food. Modern applications of geothermal energy began with the creation of the first geothermal well, by workers who accidentally drilled into one in Hungary in 1867. Today Hungary is the world's leading producer and consumer of geothermal energy, followed closely by Italy and Iceland. The word *fumarole,* used earlier in this essay, is Italian in origin, and it is said that the fumaroles near the town of Lardarello, used for the production of electricity since 1904, once inspired Dante Alighieri's (1265–1321) vision of the Inferno, as captured in his celebrated work by that name. As for Iceland, more than 99% of the buildings in the capital city of Reykjavik use geothermal energy for heat.

Heat is not the primary human application for geothermal energy. As noted earlier, in the context of the first law of thermodynamics, energy can be converted from one form to another. Thus, geothermal energy is applied for the creation of electromagnetic energy: steam or heated water from the ground runs turbines, which produce electricity.

Geothermal energy has enormous advantages, including the fact that its raw materials (heated water and steam) are free and relatively inexpensive to extract. It is also inexpensive environmentally, causing virtually no air pollution. Will geothermal energy ever significantly compete with fossil fuels as a significant source of energy for humans? It is conceivable, but at present a number of barriers exist. Geothermal resources exist only in very specific parts of the world, and the extraction of the raw materials may release noxious gases, such as hydrogen sulfide (the same compound that gives intestinal gas its smell). Also, ironically, there are environmental concerns, not because of true damage but because geothermal mines often pose a threat of sight pollution in the midst of otherwise gorgeous natural settings.

WHERE TO LEARN MORE

Altman, Linda Jacobs. *Letting Off Steam: The Story of Geothermal Energy.* Minneapolis, MN: Carolrhoda Books, 1989.

Brown, Warren. *Alternative Sources of Energy.* New York: Chelsea House, 1994.

Cox, Reg, and Neil Morris. *The Natural World.* Philadelphia: Chelsea House, 2000.

Earth's Energy Budget (Web site). <http://radar.metr.ou.edu/OK1/meteorology/EnergyBudget.html>.

Earth's Energy Budget, or Can You Spare a Sun? (Web site). <http://www.strategies.org/LESSON8.html>.

Fowler, Allan. *Energy from the Sun.* New York: Children's Press, 1997.

Gutnik, Martin J., and Natalie Browne-Gutnik. *The Energy Question: Thinking About Tomorrow.* Hillside, NJ: Enslow Publishers, 1993.

Hancock, Paul L., and Brian J. Skinner. *The Oxford Companion to the Earth.* New York: Oxford University Press, 2000.

Knapp, Brian J. *Earth Science: Discovering the Secrets of the Earth.* Illus. David Woodroffe and Julian Baker. Danbury, CT: Grolier Educational, 2000.

Skinner, Brian J., Stephen C. Porter, and Daniel B. Botkin. *The Blue Planet: An Introduction to Earth System Science.* 2nd ed. New York: John Wiley and Sons, 1999.

EARTH'S INTERIOR

EARTH'S INTERIOR

CONCEPT

For the most part, this book is concerned with geologic, geophysical, and geochemical processes that take place on or near Earth's surface. Even the essay Plate Tectonics, which takes up one of the central ideas in modern earth sciences, discusses only the lithosphere and crust but not the depths of the mantle or the core. Yet there are several good reasons to study Earth's interior, even if it is not immediately apparent why this should be the case. At first glance it would seem that activities in Earth's interior could hardly be removed further from day-to-day experience. By contrast, even the Moon seems more related to daily life. At least it is something we can see and a place to which humans have traveled; on the other hand, no human has ever seen the interior of our planet, nor is anyone likely to do so. What could Earth's interior possibly have to do with everyday life? The answer may be a bit surprising. As it turns out, many factors that sustain life itself are the result of phenomena that take place far below our feet.

HOW IT WORKS

The Core: Gravity and Density

In the essay on Planetary Science, there is a discussion of an age-old question: "Why is the earth round?" or rather "Why is Earth a sphere?" The answer, explained in more detail within the context of that essay, is that gravitational force dictates a spherical shape. As far as we know, there is no such thing as a planet or sun in the shape of a cube, because for every large object, the gravitational pull from the interior forces it to assume a more or less uniform shape. Since there is no shape more perfectly uniform than that of a sphere, this is the typical form of bodies that possess large mass.

In fact, the greater the mass, the greater the tendency toward roundness. Although they are less dense than Earth, Jupiter and Saturn are certainly more massive, and therefore they are more perfectly round. Mars and the Moon, on the other hand, are less so. Earth is not perfectly round, owing to the fact that its mass bulges at the equator because it is moving; if it were still, it would be quite round indeed, a result of the great mass at its core.

A MASSIVE CORE. As we shall see, nearly a third of Earth's mass is at its core, even though the core accounts for only about a fifth of its total volume. In other words, Earth's core is exceptionally dense, and this has several implications. First of all, the planet has a powerful gravitational pull, which not only serves to keep people and other objects rooted on the surface of the solid earth but also holds our atmosphere in place.

The gravitational attraction between any two objects is related directly to mass and inversely to the distance between them. Everything in the universe exerts some degree of gravitational pull on everything else, but unless at least one of the objects is of significant mass, the total gravitational force is negligible. The reason for this—as determined by the English mathematician and physicist Sir Isaac Newton (1642–1727)—is that gravitational force between two objects is the product of their mass divided by the distance between them and multiplied by

SIR ISAAC NEWTON *(Library of Congress.)*

an extremely small quantity known as the *gravitational constant.*

In the case of Earth, there is an extremely large amount of mass at the interior. Moreover, that mass is at a relatively short distance from objects on the planet's surface—or, to put it another way, Earth has a relatively small radius. Hence its powerful gravitational pull—one of many ways that the interior of Earth affects the overall conditions of the planet.

DENSITY OF TERRESTRIAL AND JOVIAN PLANETS. Saying that a large amount of mass is concentrated in a small area on Earth is another way of saying that the planet's interior is extremely dense. As it turns out, Earth is, in fact, the densest planet in the solar system; indeed the only other planets that come close are Mercury and Venus.

Mercury, Venus, Earth, and Mars together are designated as the terrestrial planets: bodies that are small, rocky, and dense; have relatively small amounts of gaseous elements; and are composed primarily of metals and silicates. (See the essay Minerals for more on metals as well as the extremely abundant silicates.) By contrast, the Jovian planets—Jupiter, Saturn, Uranus, and Neptune—are large, low in density, and composed primarily of gases. (Scientists know little about Pluto, which was discovered in the early twentieth century. It has a density higher than any Jovian planet, but there is little basis for classifying it as a terrestrial planet.)

Saturn, which is the least dense among the planets, has a mass only about 100 times as great as that of Earth, while its volume is almost 800 times greater. Thus its density is only about 12% of Earth's. And whereas Jovian planets, such as Saturn, are mostly gaseous and solid only in their small, dense cores, Earth is extremely solid. For a Jovian planet, there is little distinction between the "atmosphere" and the surface of the planet itself, whereas anyone who has ever jumped from a great height on Earth can attest to the sharp difference between thin air and solid ground.

Beneath that solid ground is a planetary interior composed of iron, nickel, and traces of other elements. The vast mass of the interior not only gives Earth a strong gravitational pull but also, in combination with the comparatively high speed of the planet's rotation, causes Earth to have a powerful magnetic field. Furthermore, Earth is distinguished even from most terrestrial planets (among which the Moon sometimes is counted) owing to the high degree of tectonic activity beneath its surface.

PLATE TECTONICS AND THE INTERIOR

Of all the terrestrial planets, Earth is the only one on which the processes of plate tectonics take place. Tectonism is the deformation of the lithosphere, the brittle area of Earth's interior that includes the crust and upper mantle. (We take a closer look at these regions later in this essay.) The lithosphere is characterized by large, movable segments called plates, and plate tectonics is the name both of a theory and of a specialization of tectonics, or the study of tectonism.

As a realm of study, plate tectonics deals with the large features of the lithosphere and the forces that shape them. As a theory, it explains the processes that have shaped Earth in terms of plates and their movement. This theory, discussed in detail within the Plate Tectonics essay, brings together aspects of seismic (earthquake) and volcanic activity, the structures of Earth's crust, and other phenomena to provide a unifying model of Earth's evolution. It is one of the dominant concepts in the modern earth sciences.

THE FOUR PLANETS OF THE INNER SOLAR SYSTEM, WITH THE SUN. MERCURY, VENUS, EARTH, AND MARS ARE DESIGNATED THE TERRESTRIAL PLANETS—SMALL, ROCKY, DENSE, AND COMPOSED OF METALS AND SILICATES WITH FEW GASEOUS ELEMENTS. *(© Photo Researchers. Reproduced by permission.)*

THE IMPORTANCE OF TECTONIC ACTIVITY. As discussed in Plate Tectonics, there is a difference in thickness between continental and oceanic plates on Earth. By contrast, the other terrestrial planets have crusts of fairly uniform thickness, suggesting that they have experienced little in the way of tectonic activity. Several other factors indicate that Earth is by far the most prone to tectonic activity.

Earth's core is enormous, larger than the entire planet Mercury. This means that there is a large area of high pressure and high heat driving tectonic processes, as we discuss later in this essay. In addition, Earth has a relatively thin lithosphere, meaning that the effects of heat below the lithosphere are manifested dramatically above it in the form of shifting plates and the results of such shifts—for instance, mountain building.

REAL-LIFE APPLICATIONS

"DIGGING TO CHINA"

As children, many people growing up in the West heard something along these lines: "If you could dig a hole straight through the earth, you would end up in China." This might be more or less literally true, since eastern China is on the opposite side of the planet from eastern North America. (Southeast Asia, however, is farther away, because it is more exactly opposite the eastern seaboard.) Even with the most sophisticated equipment imaginable, however, it is unlikely that anyone will put a hole straight through Earth.

The idea of "digging to China" may have raised a new question in many a child's mind. Suppose a person were to dig a hole through the Earth and jump down into it. What would happen? Gravity would carry the person to the center of the earth, but after that, would he or she just go on flying past the gravitational center of the planet? It is a good question, and the likelihood is that the powerful gravitational force at the center of the earth would hold the person there. Again, however, the likelihood of ever conducting such an experiment—for instance, with a steel ball that emitted a radio signal—is slim.

HOW DEEP? The reason for this slim likelihood can be illustrated by visiting some of the world's deepest mines. There is, for instance, the Homestake Gold Mine in South Dakota, one of the deepest mines in the United States, which extends to about 8,500 ft. (2,591 m) below the surface. This is about 1.6 mi. (2.6 km), almost six times as deep as the height of the world's tallest building, the Petronas Towers in Malaysia.

Impressive as the Homestake is, it is almost insignificant when compared with the Western Deeps Gold Mine, near Carletonville, South Africa, which reaches down about 13,000 ft. (3,962 m)! In a mine such as the Western Deeps, or even the Homestake, temperatures can reach 140°F (60°C), which makes working in such an environment extremely hazardous. Mines are air-conditioned to make them bearable, but even so, there are other dangers associated with the great depth. For instance, the pressure caused by the rocks lying above the mine may become so great that rocks in the wall shatter spontaneously.

It is no wonder, then, that workers in extremely deep gold mines and diamond mines are well paid or that their insurance premiums are very expensive. Yet even the Western Deeps is not the deepest spot where humans have drilled holes on Earth. Scientists in Sweden and Russia have overseen the drilling of deep holes purely for research purposes, while in Louisiana and Oklahoma, a few such holes have been drilled in the process of exploring for petroleum. The deepest of these holes are at Andarko Basin, Oklahoma, and the Kola Peninsula, Russia, where artificial holes extend to a staggering 7.5 mi. (12 km). This is more than three times as far down as the Western Deeps, and it is hard to imagine how any human could survive at such depths.

HOW DO WE KNOW? Even these deepest excavations represent only 0.2% of the distance from Earth's surface to its core, which is about 3,950 mi. (6,370 km) below our feet. Given that fact, one might wonder exactly how it is that earth scientists—particularly geophysicists—claim to know so much about what lies beneath the crust. In fact, they have a number of fascinating tools and methods at their disposal.

Among these tools are such rocks as kimberlite and ophiolite, which originate deep in the crust and mantle but move upward to the surface. In addition, meteorites that have landed on Earth are believed to be similar to the rocks at the mantle and core, since the planet was originally a cloud of gas around which solid materials began to form as a result of bombardment from outer space (see the essays Sun, Moon, and Earth and Planetary Science). Most important of all are seismic, or earthquake, waves, whose speed, motion, and direction tell us a great deal about the materials through which they have passed and the distances over which they have traveled.

AN IMAGINARY JOURNEY

Having established just how far humans would have to go to penetrate even just below Earth's crust with existing technology, let us now pretend that such obstacles have been overcome. In this imaginary situation, through a miracle of science let us say that there really is a hole straight through the earth and an elevator that passes through it.

This, of course, raises still more complications, aside from the gravitational problem mentioned earlier. Among other things, our elevator would have to be made of a heat-resistant material, given the temperatures we are likely to encounter in our descent. It may have sounded hot in the gold mines of South Africa and South Dakota, but that will seem cool by the time we reach Earth's core, which is as hot as the Sun's surface.

STARTING OUT. For now, however, we will throw all those logistical problems out the window and begin our journey to the center of the earth. In so doing, we will pass through three major regions—crust, mantle, and core—as well as several subsidiary realms within. By far the smallest of these is the crust, which is also the only part about which we know anything from direct experience.

Very quickly we find ourselves passing through the A, B, and C horizons of soil discussed in the Soil essay, and soon we are passing through bedrock into the main part of the crust. Bedrock might be only 5-10 ft. deep (1.5–3 m), or it might be half a mile deep (0.8 km) or perhaps even deeper. Although this is a long way for a person to dig, we still have barely scratched the surface.

As noted earlier, there is a difference between continental crust and oceanic crust. We will ignore the details here, except to say that the continental crust is thicker but the oceanic crust is denser. Thus, the continents are at a higher elevation than the oceans around them. Depending on whether the crust is oceanic or continental, we have between 3 mi. and 40 mi. (5–70 km) to travel before we begin to pass out of the crust and into the mantle.

THE LITHOSPHERE, SEISMOLOGY, AND REMOTE SENSING. The transition from crust to mantle is an abrupt one, marked by the boundary zone known as the Mohorovicic discontinuity. Sometimes called the M-discontinuity or, more commonly, the Moho, it was the discovery of the Croatian geologist Andrija Mohorovicic (1857–1936). On October 8, 1909, while studying seismic waves from an earthquake in southeastern Europe, Mohorovicic noticed that the speed of the waves increased dramatically at a depth of about 30 mi. (50 km).

Since waves travel faster through denser materials, Mohorovicic reasoned that there must be an abrupt transition from the rocky material in the Earth's crust to denser rocks below. His discovery is an excellent example of remote sensing (see Remote Sensing), whereby earth scientists are able to study places and phenomena that are impossible to observe directly.

After the Moho, which is only about 0.1-1.9 mi. (0.2–3 km) thick, we enter the mantle—or, more specifically, the lithospheric mantle. This subregion may extend to depths between 30 mi. and 60 mi. (50–100 km) and is much more dense than the crust. Like the crust, it is brittle, solid, and relatively cool compared with the regions below; hence, the crust and lithospheric mantle are lumped together as the lithosphere.

The Asthenosphere and Its Impact

At the base of the lithosphere, we pass through another transition zone, known as the Gutenberg low-velocity zone (named after the German-born American seismologist Beno Gutenberg [1889–1960]), where the speed of seismic waves again increases dramatically. After that, we enter a layer of much softer material, known as the asthenosphere. The material in the asthenosphere is soft not because it is weak—on the contrary, it is made of rock—but because it is under extraordinarily high pressure.

What happens in the asthenosphere plays a powerful role in life on the surface. The plates of the lithosphere float, as it were, atop the molten rock of the asthenosphere, which forces these plates against one another as though they were ice cubes floating in a bowl of water in constant motion. This motion is the phenomenon of plate tectonics, which, as we have discussed, quite literally shapes the world we know.

VOLCANISM AND THE ATMOSPHERE. Plate tectonics is responsible not only for such phenomena as the creation of mountains but also, by influencing the development of volcanoes, indirectly for Earth's atmosphere. In the first few billion years of the planet's existence, the action of volcanoes brought water vapor, carbon dioxide, nitrogen, sulfur, and sulfur compounds from the planet's interior to its surface. This was critical to the formation of the air we breathe today. Additionally, volcanic activity plays a significant role in the carbon cycle, whereby that vital element is circulated through various earth systems (see Biogeochemical Cycles and Carbon Cycle).

Earth and Venus stand alone among terrestrial planets as the only two still prone to volcanic activity. (By contrast, Mercury and the Moon have long been dead volcanically, and volcanism on Mars seems to have ended at some point during the past billion years.) This is significant, because even though all the planets possess more or less the same chemical elements, volcanoes are critical to distributing those elements.

In addition, volcanic activity, as well as the heat from Earth's interior that drives it and other tectonic phenomena, is an important influence on the separation of chemical compounds. When Earth formed some 4.5 billion years ago, heavier compounds—among them, those containing iron—sank toward the planet's core. At the same time, lighter ones began to rise into the atmosphere. Among these compounds was oxygen, which is clearly essential to the life of humans and other animals. This separation of compounds continues on Earth, owing to the large amount of heat that emanates from the interior.

GEOTHERMAL ENERGY. One would hardly guess that our atmosphere—or the circulation of carbon, a key component in all lifeforms—could be the indirect product of activity that takes place at least 60 mi.(100 km) below our feet. Nor is this the only illustration of the impact that Earth's interior exerts on our world. The interior of Earth is also responsible for the action that produces geothermal energy, discussed in detail within Energy and Earth.

Geothermal energy provides heating and electricity for several countries and is responsible for the dramatic effect of such phenomena as "Old Faithful" at Yellowstone Park in Wyoming. It is also the source behind the soothing natural springs found in such well-known resorts as Warm Springs, Georgia (a favorite getaway for President Franklin D. Roosevelt, who died there in 1945), and Hot Springs, Arkansas, the hometown of another president, Bill Clinton.

Mesosphere to Inner Core: Geomagnetism and Gravity

After we pass through the base of the asthenosphere, we are still only 155 mi. (250 km) deep. Now we are in the mesosphere, which extends to a depth of 1,800 mi. (2,900 km) and includes several other discontinuities, or thresholds of change. We will not discuss the details of these discontinuities here, except to note that they indicate changes in geochemical composition: for example, at 400 mi. (650 km) there appears to be a marked increase in the ratio of iron to magnesium.

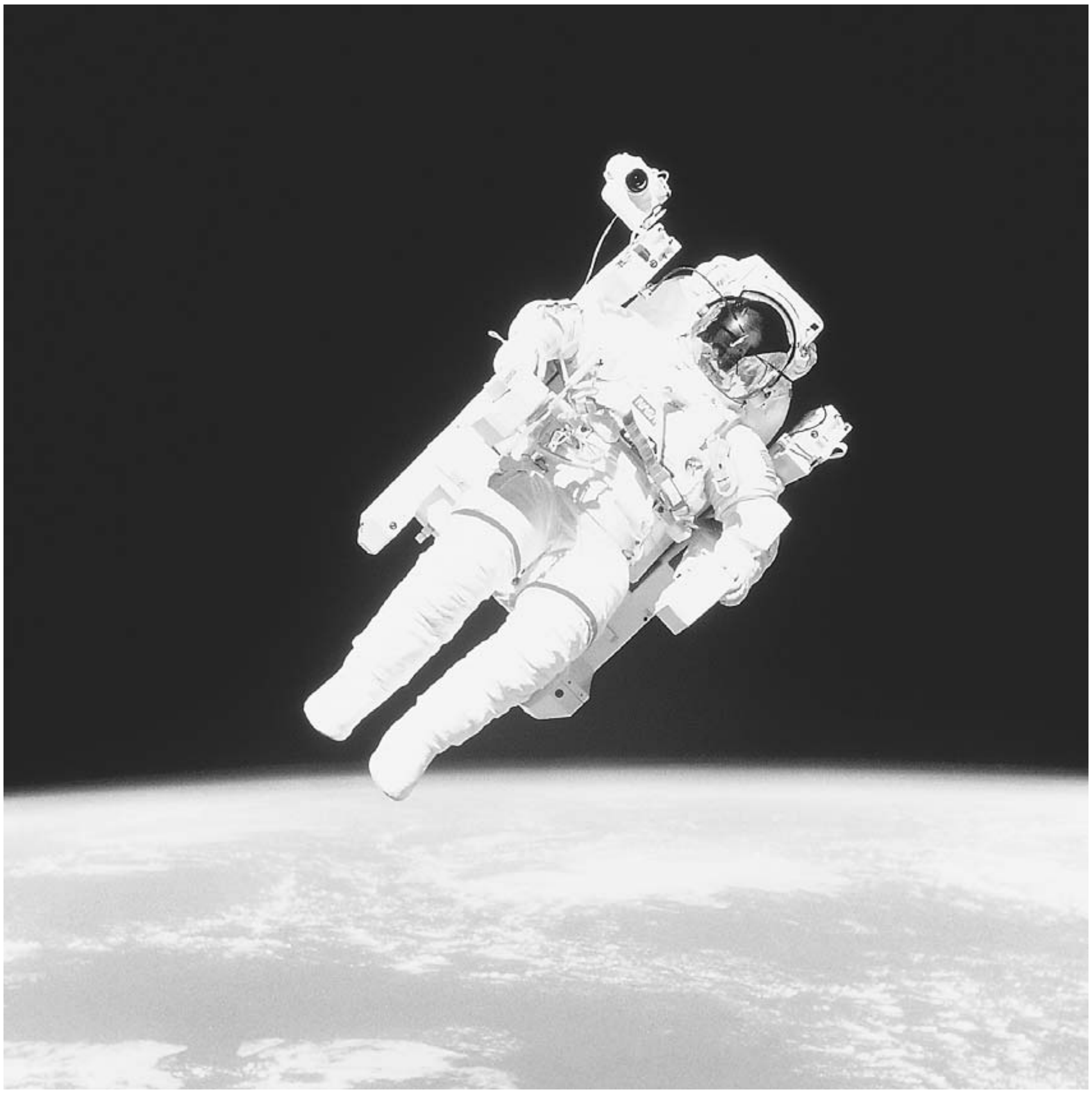
The Gutenberg discontinuity, or the core-mantle boundary (CMB), marks our entrance to the core. By now it has become very, very hot. Whereas the lithospheric mantle is about 1,600°F (870°C), the bottom of the lithosphere is about 4,000–6,700°F (2,200–3,700°C). By the time we get to the inner core, we may be confronted with temperatures as high as 13,000°F (7,200°C)—which, if this is true, would make the center of Earth about 50% hotter than the surface of the Sun!

EARTH'S FIERY HEART. At such temperatures, one would expect the rock of the outer core to be entirely molten, and indeed it is. In fact, it is this difference in phase or state of matter that marks the change from mantle (which is partially solid) to the liquid outer core. The region between the mantle and the outer core is one of undulating boundaries due to convection (see Convection), which may be the driving force behind the plate tectonic activity that occurs at a much higher level. In addition, the eddies and currents of molten iron in the core are ultimately responsible for the planet's magnetic field (see Geomagnetism).

The boundary between the mantle and core is lower today than it once was, a sign that our planet is slowly aging. If there ever comes a point when the heat is entirely dissipated, as may perhaps happen many billions of years from now, that could well be the end of Earth as a "living" planet. If we did not have a mantle and core, with all their heat, pressure, and resulting tectonic activity, Earth would be as dead as the Moon, whose interior is relatively cool.

The distinction between the outer and inner cores, which starts at a depth of about 3,150 mi. (5,100 km), comes from the fact that here, too, there is a phase change—in this case from liquid, molten material back to solid. This has nothing to do with cooling, since, as we have noted, the inner core is almost unimaginably hot; rather, it is a result of the immense pressures apparent at this depth.

GRAVITY ALWAYS WINS. It is interesting to note that the core constitutes only about 16% of the planet's volume but 32% of its mass. As we discussed near the beginning of this essay, enormous gravitational force exists between two objects when at least one of them has a relatively large amount of mass and the distance between them is great. Thus, Earth's mass, concentrated deep in its interior, helps hold our world—people, animals, plants, buildings, and so forth—in place. It also keeps our atmosphere firmly rooted as well. Without an extensive gravitational field of the kind that Earth possesses, significantly less massive bodies, such as the Moon or Mercury, have no atmosphere. In this and many another way, it turns out that life on



THE ASTRONAUT BRUCE MCCANDLESS FLOATS FREELY IN SPACE, OUTSIDE EARTH'S ATMOSPHERE, DURING A SHUTTLE MISSION. EARTH'S VAST INTERIOR MASS GIVES IT A STRONG GRAVITATIONAL PULL, HELPING ROOT THE PEOPLE AND MATERIALS OF OUR WORLD AND HOLDING OUR ATMOSPHERE IN PLACE. *(NASA. Reproduced by permission.)*

Earth's surface depends heavily on what goes on its ultra-hot, extremely pressurized interior.

A Bizarre Postscript

Given the vast amount of power in Earth's interior, it is no wonder that it has long fascinated humans—even before science possessed any sort of intelligent understanding with regard to the contents of that interior. The ancients offered all manner of fascinating speculation regarding the contents of Earth: it was hollow, some said, while others claimed that it contained one substance or another—perhaps even a heart of gold.

Such imaginative musings continued well into the Middle Ages, when the Italian poet Dante Alighieri (1265–1321) described an allegorical journey through Earth's interior in his epochal *Divine Comedy.* This epic poem depicts the inside of the planet as concentric circles of hell, descending toward the fiery core, where Satan himself resides. Beyond this lies Purgatory and further still—on the other side of Earth—Heaven, the New Jerusalem.

By the time the French writer Jules Verne (1828–1905) wrote *Journey to the Center of the Earth* almost six centuries later, scientific knowledge regarding Earth's interior had increased dramatically, though many of the significant discoveries we have examined here—for example, the Moho—still lay in the future. In any case, the

KEY TERMS

ASTHENOSPHERE: A region of extremely high pressure underlying the lithosphere, where rocks are deformed by enormous stresses. The asthenosphere lies at a depth of about 60 mi. to 215 mi. (about 100–350 km).

ATMOSPHERE: In general, an atmosphere is a blanket of gases surrounding a planet. Unless otherwise identified, however, the term refers to the atmosphere of Earth, which consists of nitrogen (78%), oxygen (21%), argon (0.93%), and other substances that include water vapor, carbon dioxide, ozone, and noble gases such as neon (0.07%).

CORE: The center of Earth, an area constituting about 16% of the planet's volume and 32% of its mass. Made primarily of iron and another, lighter element (possibly sulfur), it is divided between a solid inner core with a radius of about 760 mi. (1,220 km) and a liquid outer core about 1,750 mi. (2,820 km) thick. For terrestrial planets, in general, core refers to the center, which in most cases is probably molten metal of some kind.

CRUST: The uppermost division of the solid Earth, representing less than 1% of its volume and varying in depth from 3 mi. to 37 mi. (5–60 km). Below the crust is the mantle.

GEOCHEMISTRY: A branch of the earth sciences, combining aspects of geology and chemistry, that is concerned with the chemical properties and processes of Earth—in particular, the abundance and interaction of chemical elements and their isotopes.

GEOLOGY: The study of the solid earth, in particular, its rocks, minerals, fossils, and land formations.

GEOPHYSICS: A branch of the earth sciences that combines aspects of geology and physics. Geophysics addresses the planet's physical processes as well as its gravitational, magnetic, and electric properties and the means by which energy is transmitted through its interior.

GEOSPHERE: The upper part of Earth's continental crust, or that portion of the solid earth on which human beings live and which provides them with most of their food and natural resources.

tone of Verne's work was that of a new literary style, science fiction. Pioneered by Verne and the British writer H. G. Wells, science fiction could not have been less like Dante's poetry, infused as it was with spirituality and mystery.

SCREAMS OF THE DAMNED? In the early 1990s, an urban legend of sorts brought together the science fiction of Verne, the religious vision of Dante, and a number of other, less pleasant strains—including ignorance and, on the part of its originators, the willingness to deceive. This "urban legend" did not involve crocodiles in sewers, ghostly hitchhikers, or the other usual fodder; instead, it concerned the center of the earth—where, it was claimed, hell had been discovered.

The full account appeared on Ship of Fools (see "Where to Learn More"), a Web site operated by Rich Buhler—himself a Christian minister and a debunker of what he has called "Christian urban legends." As Buhler reported, the story gained so much support that it appeared on Trinity Broadcasting Network (TBN), a major evangelical television outlet. According to the TBN

KEY TERMS CONTINUED

JOVIAN PLANETS: The planets between Mars (the last terrestrial planet) and Pluto, all of which are large, low in density, and composed primarily of gases.

LITHOSPHERE: The upper layer of Earth's interior, including the crust and the brittle portion at the top of the mantle.

MANTLE: The thick, dense layer of rock, approximately 1,429 mi. (2,300 km) thick, between Earth's crust and its core. In reference to the other terrestrial planets, mantle simply means the area of dense rock between the crust and core.

ORGANIC: At one time chemists used the term *organic* only in reference to living things. Now the word is applied to most compounds containing carbon, with the exception of carbonates (which are minerals) and oxides, such as carbon dioxide.

PLATE TECTONICS: The name both of a theory and of a specialization of tectonics. As an area of study, plate tectonics deals with the large features of the lithosphere and the forces that shape them. As a theory, it explains the processes that have shaped Earth in terms of plates and their movement. Plate tectonics theory brings together aspects of continental drift, seafloor spreading, seismic and volcanic activity, and the structures of Earth's crust to provide a unifying model of Earth's evolution. It is one of the dominant concepts in the modern earth sciences.

PLATES: Large, movable segments of the lithosphere.

SEISMIC WAVE: A packet of energy resulting from the disturbance that accompanies a strain on rocks in the lithosphere.

SEISMOLOGY: The study of seismic waves as well as the movements and vibrations that produce them.

TECTONICS: The study of tectonism, including its causes and effects, most notably mountain building.

TECTONISM: The deformation of the lithosphere.

TERRESTRIAL PLANETS: The four inner planets of the solar system: Mercury, Venus, Earth, and Mars. These are all small, rocky, and dense; have relatively small amounts of gaseous elements; and are composed primarily of metals and silicates. Compare with Jovian planets.

report, Russian geologists had drilled a hole some 8.95 mi. (14.4 km) into Earth's crust and heard screams, which supposedly came from condemned souls in the nether regions.

As the embellished details of the story began to unfold, it turned out that the Russian geologists had found the temperatures to be much higher than expected: 2,000°F (1,093°C). Also, their drilling had unleashed a bat that flew out of hell with the words "I have conquered" inscribed in Russian on its wings. Buhler and his team traced this bizarre tale to Finland and then back to southern California. As to how the story originated, Buhler noted the drilling at the Kola Peninsula, which we mentioned earlier in this essay. The depth cited in the rumor, however, was greater than that which the drilling at Kola reached, and the temperatures claimed were much higher than what one actually would encounter at that depth.

"It is possible that somewhere in the world there has been a spooky experience during deep drilling operations," Buhler concluded. Nonetheless, "characteristic of many urban legends, this story was alleged to have occurred in an obscure part of the world where it would be virtually impossible to track down the facts. And once the story got started, people began quoting one another's newsletters to validate their own. This is the stuff of which tabloid newspapers are made." In the end, the "screams of hell" offered nothing of value in terms of either science or religion, but it proved to be an excellent example of human beings' fascination with, and latent terror of, Earth's interior.

WHERE TO LEARN MORE

Buhler, Rich. "Drilling for Hell," *Ship of Fools* (Web site). <http://ship-of-fools.com/Myths/03Myth.html#Top>.

Darling, David J. *Could You Ever Dig a Hole to China?* Minneapolis, MN: Dillon Press, 1990.

De Bremaecker, Jean-Claude. *Geophysics: The Earth's Interior.* New York: John Wiley and Sons, 1985.

Earth's Interior (Web site). <http://www.seismo.unr.edu/ftp/pub/louie/class/100/interior.html>.

Earth's Interior and Plate Tectonics (Web site). <http://www.star.le.ac.uk/edu/solar/earthint.html>.

Earth's Interior and Plate Tectonics (Web site). <http://www.solarviews.com/eng/earthint.htm>.

Hancock, Paul L., and Brian J. Skinner. *The Oxford Companion to the Earth.* New York: Oxford University Press, 2000.

Physical Geology: Interior of the Earth (Web site). <http://www.uh.edu/~jbutler/physical/chapter19.html>.

Rockdoctors Guide: Earth's Interior (Web site). <http://www.cobweb.net/~bug2/rock7.htm>.

Skinner, Brian J., Stephen C. Porter, and Daniel B. Botkin. *The Blue Planet: An Introduction to Earth System Science.* 2d ed. New York: John Wiley and Sons, 1999.

PLATE TECTONICS

CONCEPT

The earth beneath our feet is not dead; it is constantly moving, driven by forces deep in its core. Nor is the planet's crust all of one piece; it is composed of numerous plates, which are moving steadily in relation to one another. This movement is responsible for all manner of phenomena, including earthquakes, volcanoes, and the formation of mountains. All these ideas, and many more, are encompassed in the concept of plate tectonics, which is the name for a branch of geologic and geophysical study and for a powerful theory that unites a vast array of ideas. Plate tectonics works hand in hand with several other striking concepts and discoveries, including continental drift and the many changes in Earth's magnetic field that have taken place over its history. No wonder, then, that this idea, developed in the 1960s but based on years of research that preceded that era, is described as "the unifying theory of geology."

HOW IT WORKS

Tectonics and Tectonism

The lithosphere is the upper layer of Earth's interior, including the crust and the brittle portion at the top of the mantle. Tectonism is the deformation of the lithosphere, and the term tectonics refers to the study of this deformation, including its causes and effects, most notably mountain building. This deformation is the result of the release and redistribution of energy from Earth's core.

The interior of Earth itself is divided into three major sections: the crust, mantle, and core. The first is the uppermost division of the solid earth, representing less than 1% of its volume and varying in depth from 3 mi. to 37 mi. (5–60 km). Below the crust is the mantle, a thick, dense layer of rock approximately 1,429 mi. (2,300 km) thick. The core itself is even more dense, as illustrated by the fact that it constitutes about 16% of the planet's volume and 32% of its mass. Composed primarily of iron and another, lighter element (possibly sulfur), it is divided between a solid inner core with a radius of about 760 mi. (1,220 km) and a liquid outer core about 1,750 mi. (2,820 km) thick.

Tectonism results from the release and redistribution of energy from Earth's interior. There are two components of this energy: gravity, a function of the enormous mass at the core, and heat from radioactive decay. (For more about gravity, see Gravity and Geodesy. The heat from Earth's core, the source of geothermal energy, is discussed in Energy and Earth.) Differences in mass and heat within the planet's interior, known as pressure gradients, result in the deformation of rocks.

DEFORMATION OF ROCKS. Any attempt to deform an object is referred to as stress, and stress takes many forms, including tension, compression, and shear. Tension acts to stretch a material, whereas compression—a type of stress produced by the action of equal and opposite forces, whose effect is to reduce the length of a material—has the opposite result. (Compression is a form of pressure.) As for shear, this is a kind of stress resulting from equal and opposite forces that do not act along the same line. If a thick, hardbound book is lying flat and one pushes the front cover from the side so that the covers and pages are no longer perfectly aligned, this is an example of shear.

Under the effects of these stresses, rocks may bend, warp, slide, or break. They may even flow, as though they were liquids, or melt and thus truly become liquid. As a result, Earth's interior may manifest faults, or fractures in rocks, as well as folds, or bends in the rock structure. The effects of this activity can be seen on the surface in the form of subsidence, which is a depression in the crust, or uplift, which is the raising of crustal materials. Earthquakes and volcanic eruptions also may result.

There are two basic types of tectonism: orogenesis and epeirogenesis. *Orogenesis* is taken from the Greek words *oros* ("mountain") and *genesis* ("origin") and involves the formation of mountain ranges by means of folding, faulting, and volcanic activity. The Greek word *epeiros* means "mainland," and epeirogenesis takes the form of either uplift or subsidence. Of principal concern in the theory of plate tectonics, as we shall see, is orogenesis, which involves more lateral, as opposed to vertical, movement.

CONTINENTAL DRIFT

If one studies a world map for a period of time, one may notice something interesting about the shape of Africa's west coast and that of South America's east coast: they seem to fit together like pieces of a jigsaw puzzle. Early in the twentieth century, two American geologists, Frank Bursley Taylor (1860–1938) and Howard Baker, were among the first scientists to point out this fact. According to Taylor and Baker, Europe, the Americas, and Africa all had been joined at one time. This was an early version of continental drift, a theory concerning the movement of Earth's continents.

Continental drift is based on the idea that the configuration of continents was once different than it is today, that some of the individual landmasses of today once were joined in other continental forms, and that the landmasses later moved to their present locations. Though Taylor and Baker were early proponents, the theory is associated most closely with the German geophysicist and meteorologist Alfred Wegener (1880–1930), who made the case for continental drift in *The Origin of Continents and Oceans* (1915).

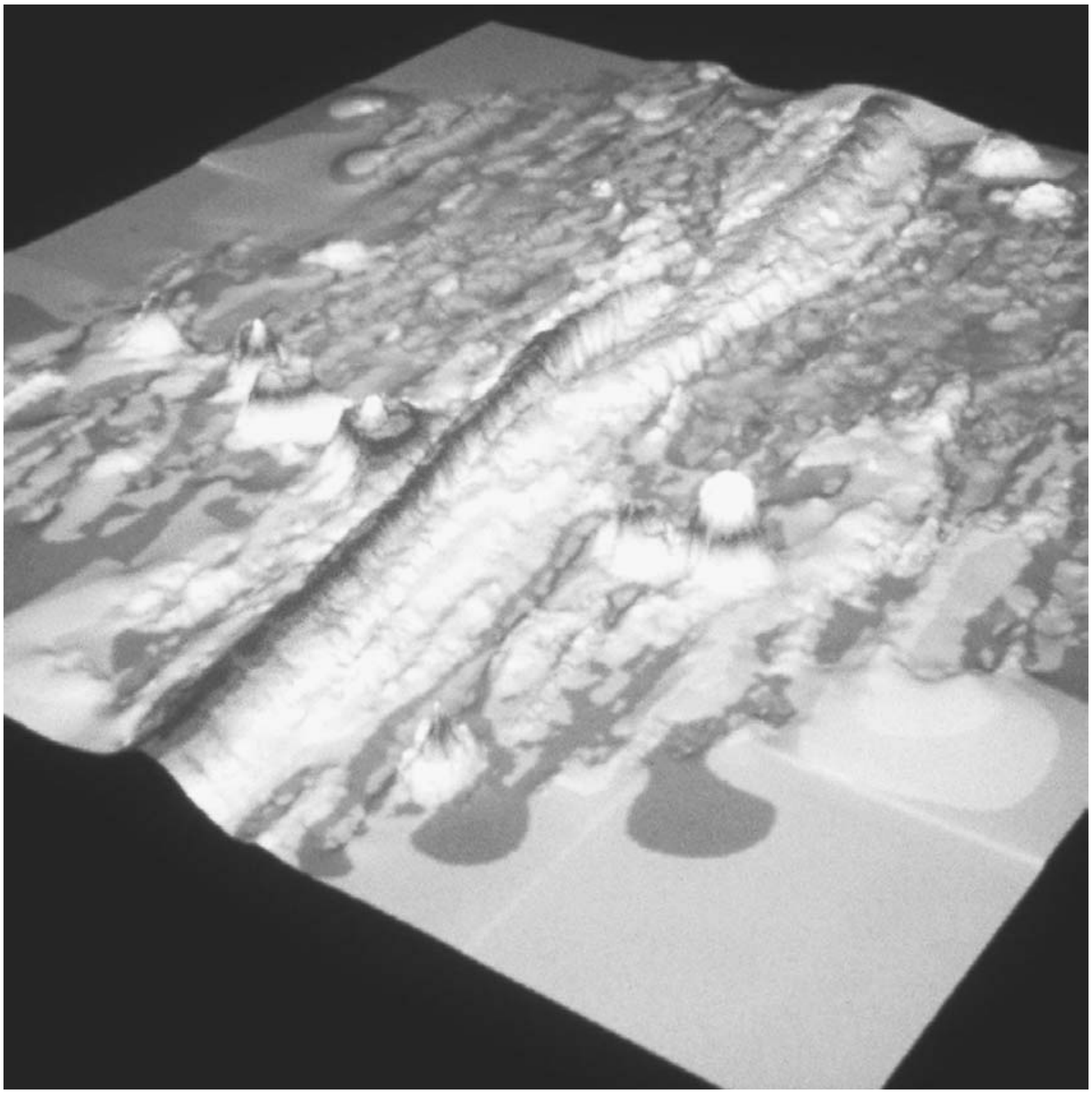
PANGAEA, LAURASIA, AND GONDWANALAND. According to Wegener, the continents of today once formed a single supercontinent called Pangaea, from the Greek words *pan* ("all") and *gaea* ("Earth"). Eventually, Pangaea split into two halves, with the northern continent of Laurasia and the southern continent of Gondwanaland, sometimes called Gondwana, separated by the Tethys Sea. In time, Laurasia split to form North America, the Eurasian landmass with the exception of the Indian subcontinent, and Greenland. Gondwanaland also split, forming the major southern landmasses of the world: Africa, South America, Antarctica, Australia, and India.

The Austrian geologist Eduard Suess (1831–1914) and the South African geologist Alexander du Toit (1878–1948), each of whom contributed significantly to continental drift theory, were responsible for the naming of Gondwanaland and Laurasia, respectively. Suess preceded Wegener by many years with his theory of Gondwanaland, named after the Gondwana region of southern India. There he found examples of a fern that, in fossilized form, had been found in all the modern-day constituents of the proposed former continent. Du Toit, Wegener's contemporary, was influenced by continental drift theory and improved on it greatly.

FORMATION OF THE CONTINENTS. Today continental drift theory is accepted widely, in large part owing to the development of plate tectonics, "the unifying theory in geology." We examine the evidence for continental drift, the arguments against it, and the eventual triumph of plate tectonics in the course of this essay. Before going on, however, let us consider briefly the now-accepted timeline of events described by Wegener and others.

About 1,100 million years ago (earth scientists typically abbreviate this by using the notation 1,100 Ma), there was a supercontinent named Rodinia, which predated Pangaea. It split into Laurasia and Gondwanaland, which moved to the northern and southern extremes of the planet, respectively. Starting at about 514 Ma, Laurasia drifted southward until it crashed into Gondwanaland about 425 Ma. Pangaea, surrounded by a vast ocean called Panthalassa ("All Ocean"), formed approximately 356 Ma.

In the course of Pangaea's formation, what is now North America smashed into northwestern



A map of part of the East Pacific Rise, a mid-ocean ridge to the west of Central America that marks the boundary between the Pacific and Cocos tectonic plates. *(© Dr. Ken MacDonald/Photo Researchers. Reproduced by permission.)*

Africa, forming a vast mountain range. Traces of these mountains still can be found on a belt stretching from the southern United States to northern Europe, including the Appalachians. As Pangaea drifted northward and smashed into the ocean floor of Panthalassa, it formed a series of mountain ranges from Alaska to southern South America, including the Rockies and Andes. By about 200 Ma, Pangaea began to break apart, forming a valley that became the Atlantic Ocean. But the separation of the continents was not a "neat" process: today a piece of Gondwanaland lies sunken beneath the eastern United States, far from the other landmasses to which it once was joined.

By about 152 Ma, in the late Jurassic period, the continents as we know them today began to take shape. By about 65 Ma, all the present continents and oceans had been formed for the most part, and India was drifting north, eventually smashing into southern Asia to shape the world's tallest mountains, the Himalayas, the Karakoram Range, and the Hindu Kush. This process is not finished, however, and geologists believe that some 250–300 million years from now, Pangaea will re-form.

EVIDENCE AND ARGUMENTS. As proof of his theory, Wegener cited a wide variety of examples, including the apparent fit between the coastlines of South America and western Africa as well as that of North America and northwestern Africa. He also noted the existence of rocks apparently gouged by glaciers in southern Africa, South America, and India, far from modern-day glacial activity. Fossils in South America matched those in Africa and Australia, as Suess had observed. There were also signs that mountain ranges continued between continents—not only those apparently linking North America and Europe but also ranges that seemed to extend from Argentina to South Africa and Australia.

By measurements conducted over a period of years, Wegener even showed that Greenland was drifting slowly away from Europe, yet his theory met with scorn from the geoscience community of his day. If continents could plow through oceanic rock, some geologists maintained, then they would force up mountains so high that Earth would become imbalanced. As for his claim that matching fossils in widely separated regions confirmed his theory of continental drift, geologists claimed that this could be explained by the existence of land bridges, now sunken, that once had linked those areas. The apparent fit between present-day landmasses could be explained away as coincidence or perhaps as evidence that Earth simply was expanding, with the continents moving away from one another as the planet grew.

Introduction to Plate Tectonics

Though Wegener was right, as it turned out, his theory had one major shortcoming: it provided no explanation of exactly *how* continental drift had occurred. Even if geologists had accepted his claim that the continents are moving, it raised more questions than answers. A continent is a very large thing simply to float away; even an aircraft carrier, which is many millions of times lighter, has to weigh less than the water it displaces, or it would sink like a stone. In any case, Wegener never claimed that continents floated. How, then, did they move?

The answer is plate tectonics, the name both of a theory and of a specialization of tectonics. As an area of study, plate tectonics deals with the large features of the lithosphere and the forces that fashion them. As a theory, it explains the processes that have shaped Earth in terms of plates (large movable segments of the lithosphere) and their movement. Plate tectonics theory brings together aspects of continental drift, seafloor spreading (discussed later), seismic and volcanic activity, and the structures of Earth's crust to provide a unifying model of Earth's evolution.

It is hard to overemphasize the importance of plate tectonics in the modern earth sciences; hence, its characterization as the "unifying theory." Its significance is demonstrated by its inclusion in the book *The Five Biggest Ideas of Science,* cited in the bibliography for this essay. Alongside plate tectonics theory in that volume are four towering concepts of extraordinary intellectual power: the atomic model, or the concept that matter is made up of atoms; the periodic law, which explains the chemical elements; big bang theory, astronomers' explanation of the origins of the universe (see Planetary Science); and the theory of evolution in the biological sciences.

THE PIECES COME TOGETHER. In 1962 the United States geologist Harry Hammond Hess (1906–1969) introduced a new concept that would prove pivotal to the theory of plate tectonics: seafloor spreading, the idea that seafloors crack open along the crest of mid-ocean ridges and that new seafloor forms in those areas. (Another American geologist, Robert S. Dietz [1914–1995], had published his own theory of seafloor spreading a year before Hess's, but Hess apparently developed his ideas first.) According to Hess, a new floor forms when molten rock called magma rises up from the asthenosphere, a region of extremely high pressure underlying the lithosphere, where rocks are deformed by enormous stresses. The magma wells up through a crack in a ridge, runs down the sides, and solidifies to form a new floor.

Three years later, the Canadian geologist John Tuzo Wilson (1908–1993) coined the term plates to describe the pieces that make up Earth's rigid surface. Separated either by the mid-ocean rifts identified earlier by Heezen or by mountain chains, the plates move with respect to one another. Wilson presented a model for their behavior and established a global pattern of faults, a sort of map depicting the movable plates. The pieces of a new theory were forming (an apt metaphor in this instance!), but as yet it had no name.

That name appeared in 1967, when D. P. Mackenzie of England and R. L. Parker of the

United States introduced the term plate tectonics. They maintained that the surface of Earth is divided into six major as well as seven minor movable plates and compared the continents to enormous icebergs—much as Wegener had described them half a century earlier. Subsequent geologic research has indicated that there may be as many as nine major plates and as many as 12 minor ones.

To test these emerging ideas, the U.S. National Science Foundation authorized a research voyage by the vessel *Glomar Challenger* in 1968. On their first cruise, through the Gulf of Mexico and the Atlantic, the *Challenger*'s scientific team collected sediment, fossil, and crust samples that confirmed the basics of seafloor spreading theory. These results led to new questions regarding the reactions between rocks and the heated water surrounding them, spawning new research and necessitating additional voyages. In the years that followed, the *Challenger* made more and more cruises, its scientific teams collecting a wealth of evidence for the emerging theory of plate tectonics.

REAL-LIFE APPLICATIONS

Early Evidence of Plate Tectonics

No single person has been as central to plate tectonics as Wegener was to continental drift or as the English naturalist Charles Darwin (1809–1882) was to evolution. The roots of plate tectonics lie partly in the observations of Wegener and other proponents of continental drift as well as in several discoveries and observations that began to gather force in the third quarter of the twentieth century.

During World War II, submarine warfare necessitated the development of new navigational technology known as sonar (*SO*und *N*avigation *A*nd *R*anging). Sonar functions much like radar (see Remote Sensing), but instead of using electromagnetic waves, it utilizes ultrasonic, or high-frequency, sound waves projected through water. Sonar made it feasible for geologists to study deep ocean basins after the war, making it possible for the first time in history to map and take samples from large areas beneath the seas. These findings raised many questions, particularly concerning the vast elevation differences beneath the seas.

EWING AND THE MOUNTAINS UNDER THE OCEAN. One of the first earth scientists to notice the curious aspects of underwater geology was the American geologist William Maurice Ewing (1906–1974), who began his work long before the war. He had gained his first experience in a very practical way during the 1920s, as a doctoral student putting himself through school. Working summers with oil exploration teams in the Gulf of Mexico off the coast of Texas had given him a basic understanding of the subject, and in the following decade he went to work exploring the structure of the Atlantic continental shelf and ocean basins.

His work there revealed extremely thick sediments covering what appeared to be high mountainous regions. These findings sharply contradicted earlier ideas about the ocean floor, which depicted it as a flat, featureless plain rather like the sandy-bottomed beaches found in resort areas. Instead, the topography at the bottom of the ocean turned out to be at least as diverse as that of the land above sea level.

HEEZEN AND THE RIFT VALLEY. During the 1950s, a team led by another American geologist, Bruce Charles Heezen (1924–1977), worked on developing an overall picture of the ocean basin's topography. Earlier work had identified a mountain range running the length of the Atlantic, but Heezen's team discovered a deep valley down the middle of the chain, running parallel to it. They described it as a rift valley, a long trough bounded by two or more faults, and compared it to a similar valley in eastern Africa.

Around the same time, a group of transatlantic telephone companies asked Heezen to locate areas of possible seismic or earthquake activity in the Atlantic. Phone company officials reasoned that if they could find the areas most likely to experience seismic activity, they could avoid placing their cables in those areas. As it turned out, earthquakes tended to occur in exactly the same region that Heezen and his team had identified as the rift valley.

The Plates and Their Interactions

The most significant plates that make up Earth's surface are as follows:

Selected Major Plates

- North American (almost all of North America and Mexico, along with Greenland and the northwestern quadrant of the Atlantic)
- South American (all of South America and the southwestern quadrant of the Atlantic)
- African (Africa, the southeastern Atlantic, and part of the Indian Ocean)
- Eurasian (Europe and Asia, excluding the Indian subcontinent, along with surrounding ocean areas)
- Indo-Australian (India, much of the Indian Ocean, Australia, and parts of the Indonesian archipelago and New Zealand)
- Antarctic (Antarctica and the Antarctic Ocean)

In addition to these plates, there are several plates that while they are designated as "major" are much smaller: the Philippine, Arabian, Caribbean, Nazca (off the west coast of South America), Cocos (off the west coast of Mexico), and Juan de Fuca (extreme western North America). Japan, one of the most earthquake-prone nations in the world, lies at the nexus of the Philippine, Eurasian, and Pacific plates.

MOVEMENT OF THE PLATES. One of the key principles of geology, discussed elsewhere in this book, is uniformitarianism: the idea that processes occurring now also occurred in the past. The reverse usually is also true; thus, as we have noted, the plates are still moving, just as they have done for millions of years. Thanks to satellite remote sensing, geologists are able to measure this rate of movement. (See Remote Sensing for more on this subject.) Not surprisingly, its pace befits the timescale of geologic, as opposed to human, processes: the fastest-moving plates are careening forward at a breathtaking speed of 4 in. (10 cm) per year. The ground beneath Americans' feet (assuming they live in the continental United States, east of the Juan de Fuca) is drifting at the rate of 1.2 in. (3 cm) every year, which means that in a hundred years it will have shifted 10 ft. (3 m).

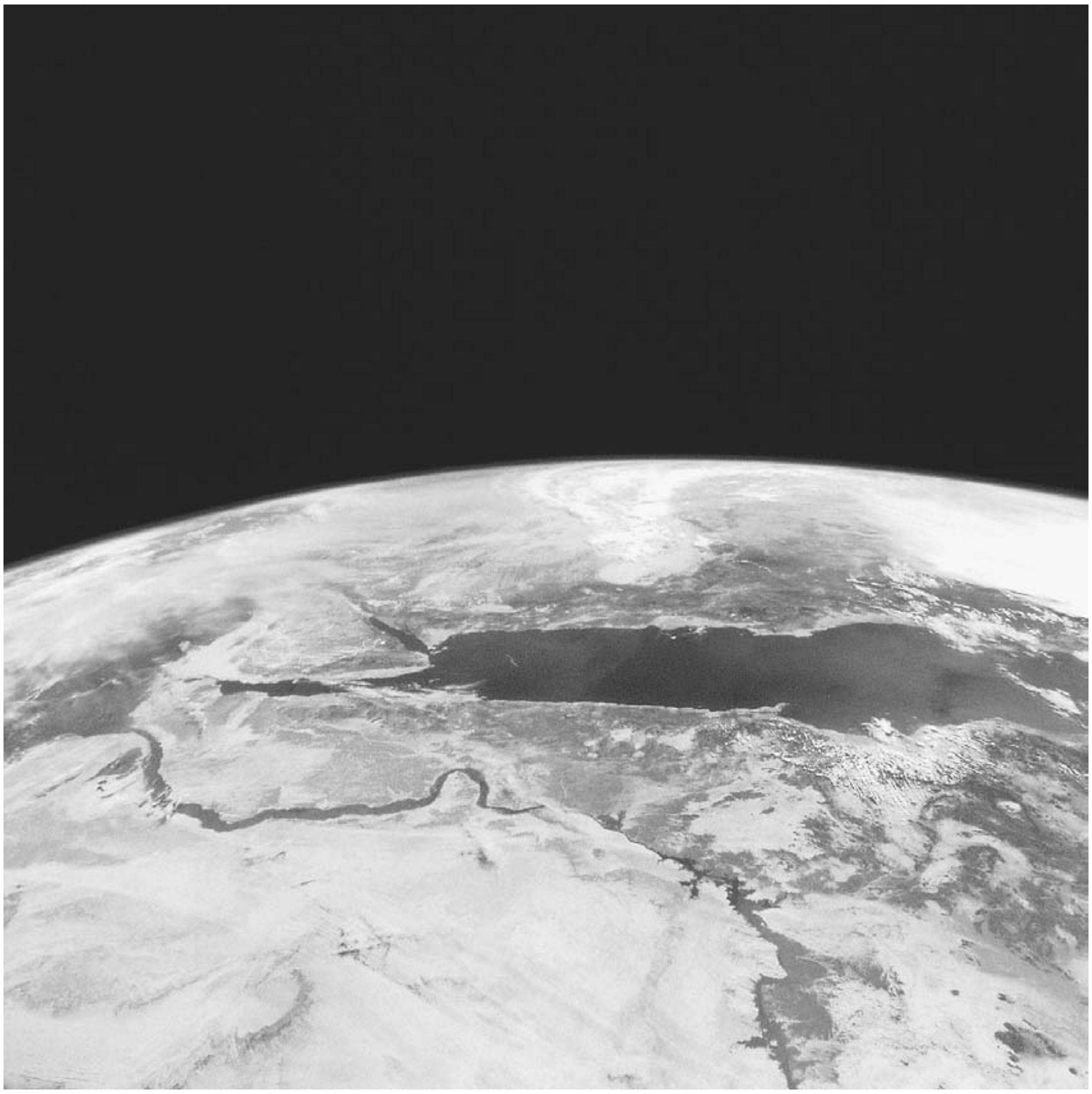
WHEN PLATES INTERACT. Plates interact by moving toward each other (convergence), away from each other (divergence), or past each other (transform motion). Convergence usually is associated with subduction, meaning that one plate is forced down into the mantle and eventually undergoes partial melting. This typically occurs in the ocean, creating a depression known as an oceanic trench. Divergence results in the separation of plates and most often is associated either with seafloor spreading or the formation of rift valleys.

There are three types of plate margins, or boundaries between plates, depending on the two types of crusts that are interacting: oceanic with oceanic, continental with continental, or continental with oceanic. The rift valleys of the Atlantic are an example of an oceanic margin where divergence has occurred, while oceanic convergence is illustrated by a striking example in the Pacific. There, subduction of the Philippine Plate by the Pacific Plate has created the Mariana Trench, which at 36,198 ft. (10,911 m) is the deepest depression on Earth.

When continental plates converge, neither plate subducts; rather, they struggle against each other like two warriors in a fight to the death, buckling, folding, and faulting to create huge mountain ranges. The convergence of the Indo-Australian and Eurasian plates has created the *highest* spots on Earth, in the Himalayas, where Mount Everest (on the Nepal-Indian border) rises to 29,028 ft. (8,848 m). Continental plates also may experience divergence, resulting in the formation of seas. An example is the Red Sea, formed by the divergence of the African and Arabian plates.

Given these facts about the interactions of oceanic and continental plates with each other, what occurs when continental plates meet oceanic ones is no surprise. In this situation, the oceanic plate meeting the continental plate is like a high-school football player squaring off against a National Football League pro tackle. It is no match: the oceanic plate easily subducts. This leads to the formation of a chain of volcanoes along the continental crust, examples being the Cascade Range in the U.S. Pacific Northwest (Juan de Fuca and Pacific plates) or the Andes (South American and Nazca plates).

Transform margins may occur with any combination of oceanic or continental plates and result in the formation of faults and earthquake zones. Where the North American Plate slides against the Pacific Plate along the California coast, it has formed the San Andreas Fault, the source of numerous earthquakes, such as the dramatic San Francisco quakes of 1906 and 1989 and the Los Angeles quake of 1994.



SHUTTLE PHOTOGRAPH OF EASTERN EGYPT SHOWS THE RED SEA AT THE TOP, WITH THE GULF OF SUEZ CONNECTING IT TO THE MEDITERRANEAN SEA. THE RED SEA WAS FORMED BY THE DIVERGENCE OF THE AFRICAN AND ARABIAN PLATES. *(© NASA/Photo Researchers. Reproduced by permission.)*

PALEOMAGNETISM

As noted earlier, plate tectonics brings together numerous areas of study in the geologic sciences that developed independently but which came to be seen as having similar roots and explanations. Among these disciplines is paleomagnetism, an area of historical geology devoted to studying the direction and intensity of magnetic fields in the past, as discerned from the residual magnetization of rocks.

Earth has a complex magnetic field whose principal source appears to be the molten iron of the outer core. In fact, the entire planet is like a giant bar magnet, with a north pole and a south pole. It is for this reason that the magnetized material in a compass points north; however, Earth's magnetic north pole is not the same as its geographic north pole. It so happens that magnetic north lies in more or less the same direction as geographic north, but as geologists in the mid–nineteenth century discovered, this has not always been the case. (For more about magnetic north and other specifics of Earth's magnetic field, see Geomagnetism.)

In 1849 the French physicist Achilles Delesse (1817–1881) observed that magnetic minerals tend to line up with the planet's magnetic field, pointing north as though they were compass needles. Nearly 60 years later, however, another French physicist, Bernard Brunhes (1867–1910),

KEY TERMS

ASTHENOSPHERE: A region of extremely high pressure underlying the lithosphere, where rocks are deformed by enormous stresses. The asthenosphere lies at a depth of about 60-215 mi. (about 100–350 km).

COMPRESSION: A form of stress produced by the action of equal and opposite forces, the effect of which is to reduce the length of a material. Compression is a form of pressure.

CONTINENTAL DRIFT: The theory that the configuration of Earth's continents was once different than it is today; that some of the individual landmasses of today once were joined in other continental forms; and that these landmasses later separated and moved to their present locations.

CONVERGENCE: A tectonic process whereby plates move toward each other. Usually associated with subduction, convergence typically occurs in the ocean, creating an oceanic trench. It is one of the three ways, along with divergence and transform motion, that plates interact.

CORE: The center of Earth, an area constituting about 16% of the planet's volume and 32% of its mass. Made primarily of iron and another, lighter element (possibly sulfur), it is divided between a solid inner core with a radius of about 760 mi. (1,220 km) and a liquid outer core about 1,750 mi. (2,820 km) thick.

CRUST: The uppermost division of the solid earth, representing less than 1% of its volume and varying in depth from 3-37 mi. (5–60 km). Below the crust is the mantle.

DIVERGENCE: A tectonic process whereby plates move away from each other. Divergence results in the separation of plates and is associated most often either with seafloor spreading or with the formation of rift valleys. It is one of the three ways, along with convergence and transform motion, that plates interact.

EPEIROGENESIS: One of two principal forms of tectonism, the other being orogenesis. Derived from the Greek words *epeiros* ("mainland") and *genesis* ("origins"), epeirogenesis takes the form of either uplift or subsidence.

FAULT: An area of fracturing between rocks resulting from stress.

FOLD: An area of rock that has been bent by stress.

GEOPHYSICS: A branch of the earth sciences that combines aspects of geology and physics. Geophysics addresses the planet's physical processes as well as its gravitational, magnetic, and electric properties, and the means by which energy is transmitted through its interior.

HISTORICAL GEOLOGY: The study of Earth's physical history. Historical geology is one of two principal branches of geology, the other being physical geology.

LITHOSPHERE: The upper layer of Earth's interior, including the crust and the brittle portion at the top of the mantle.

MA: An abbreviation used by earth scientists, meaning "million years." When an event is designated as, for instance, 160 Ma, it means that it happened 160 million years ago.

MANTLE: The thick, dense layer of rock, approximately 1,429 mi. (2,300 km) thick, between Earth's crust and its core.

KEY TERMS CONTINUED

MID-OCEAN RIDGES: Submarine mountain ridges where new seafloor is created by seafloor spreading.

OCEANIC TRENCH: A deep depression in the ocean floor caused by the convergence of plates and the resulting subduction of one plate.

OROGENESIS: One of two principal forms of tectonism, the other being epeirogenesis. Derived from the Greek words *oros* ("mountain") and *genesis* ("origin"), orogenesis involves the formation of mountain ranges by means of folding, faulting, and volcanic activity. The processes of orogenesis play a major role in plate tectonics.

PALEOMAGNETISM: An area of historical geology devoted to studying the direction and intensity of magnetic fields in the past, as discerned from the residual magnetization of rocks.

PLATE MARGINS: Boundaries between plates.

PLATE TECTONICS: The name both of a theory and of a specialization of tectonics. As an area of study, plate tectonics deals with the large features of the lithosphere and the forces that fashion them. As a theory, it explains the processes that have shaped Earth in terms of plates and their movement. Plate tectonics theory brings together aspects of continental drift, seafloor spreading, seismic and volcanic activity, and the structures of Earth's crust to provide a unifying model of Earth's evolution. It is one of the dominant concepts in the modern earth sciences.

PLATES: Large movable segments of the lithosphere.

RADIOACTIVITY: A term describing a phenomenon whereby certain materials are subject to a form of decay brought about by the emission of high-energy particles or radiation. Forms of particles or energy include alpha particles (positively charged helium nuclei), beta particles (either electrons or subatomic particles called *positrons*), or gamma rays, which occupy the highest energy level in the electromagnetic spectrum.

REMOTE SENSING: The gathering of data without actual contact with the materials or objects being studied.

RIFT: A split between two bodies (for example, two plates) that once were joined.

RIFT VALLEY: A long trough bounded by two or more faults.

SEAFLOOR SPREADING: The theory that seafloors crack open along the crests of mid-ocean ridges and that new seafloor forms in those areas.

SHEAR: A form of stress resulting from equal and opposite forces that do not act along the same line. If a thick, hardbound book is lying flat and one pushes the front cover from the side so that the covers and pages are no longer perfectly aligned, this is an example of shear.

STRESS: In general terms, any attempt to deform a solid. Types of stress include tension, compression, and shear. More specifically, stress is the ratio of force to unit area F/A, where F is force and A area.

SUBDUCTION: A tectonic process that results when plates converge and one plate forces the other down into Earth's mantle. As a result, the subducted plate eventually undergoes partial melting.

KEY TERMS CONTINUED

SUBSIDENCE: A depression in Earth's crust.

TECTONICS: The study of tectonism, including its causes and effects, most notably mountain building.

TECTONISM: The deformation of the lithosphere.

TENSION: A form of stress produced by a force that acts to stretch a material.

THEORY: A general statement derived from a hypothesis that has withstood sufficient testing.

TRANSFORM MOTION: A tectonic process whereby plates slide past each other. It is one of the three ways, along with convergence and divergence, that plates interact.

noted that in some rocks magnetic materials point *south*. This suggested one of two possibilities: either the planet's magnetic field had reversed itself over time, or the ground containing the magnetized rocks had moved. Both explanations must have seemed far-fetched at the time, but as it turned out, both are correct.

Earth's magnetic field has shifted, meaning that the magnetic north and south poles have changed places many times over the eons. In addition, the magnetic poles have wandered around the southern and northern portions of the globe: for instance, whereas magnetic north today lies in the frozen islands to the north of Canada, at about 300 Ma it was located in eastern Siberia. The movement of magnetic rocks on Earth's surface, however, has turned out to be too great to be explained either by magnetic shifts or by regional wandering of the poles. This is where plate tectonics and paleomagnetism come together.

CONFIRMATION OF PLATE TECTONIC THEORY. Rocks in Alaska have magnetic materials aligned in such a way that they once must have been at or near the equator. In addition, the orientation of magnetic materials on South America's east coast shows an affinity with that of similar materials on the west coast of Africa. In both cases, continental drift, with its driving mechanism of plate tectonics, seems the only reasonable explanation.

Thus, paleomagnetic studies have served to confirm the ideas of continental drift and plate tectonics, while research conducted at sea bolsters seafloor spreading theory. Using devices called magnetometers, geologists have found that the orientation of magnetic minerals on one side of a rift mirrors that of materials on the other side. This suggests that the new rock on either side of the rift was formed simultaneously, as seafloor spreading theory indicates.

EARTHQUAKES AND VOLCANOES

Several findings relating to earthquakes and volcanic activity also can be explained by plate tectonics. If one follows news stories of earthquakes, one may begin to wonder why such places as California or Japan have so many quakes, whereas the northeastern United States or western Europe have so few. The fact is that earthquakes occur along belts, and the vast majority of these belts coincide with the boundaries between Earth's major tectonic plates.

The same is true of volcanoes, and it is no mistake that places famous for earthquakes—the Philippines, say, or Italy—often also are known for their volcanoes. Although they are located near the center of the Pacific Plate, the islands of Hawaii are subject to plate movement, which has helped generate the volcanoes that gave those islands their origin. At the southern end of the island chain, many volcanoes are still active, while those at the northern end tend to be dormant. The reason is that the Pacific Plate as a whole is moving northward over a stationary lava source in the mantle below Hawaii. The southern islands remain poised above that source, while the northern islands have moved away from it.

The Oceanic and Continental Crusts

Given what we have seen about continental drift and seafloor spreading, it should come as no surprise to learn that, generally speaking, the deeper one goes in the ocean, the newer the crust. Specifically, the crust is youngest near the center of ocean basins and particularly along mid-ocean ridges, or submarine mountain ridges where new seafloor is created by seafloor spreading.

It also should not be surprising to learn that oceanic and continental crusts differ both in thickness and in composition. Basalt, an igneous rock (rock formed from the cooling of magma), makes up the preponderance of ocean crust, whereas much of the continental crust is made up of granite, another variety of igneous rock. Whereas the ocean crust is thin, generally 3–6 mi. (5–10 km) in depth, the continental crust ranges in thickness from 12.5–55 mi. (20–90 km). This results in a difference in thickness for the lithosphere, which is only about 60 mi. (100 km) thick beneath the oceans but about 2.5 times as thick—150 mi. (250 km)—under the continents.

WHERE TO LEARN MORE

Erickson, Jon. *Plate Tectonics: Unraveling the Mysteries of the Earth.* New York: Facts on File, 1992.

Gallant, Roy A. *Dance of the Continents.* New York: Benchmark Books, 2000.

Geology: Plate Tectonics (Web site). <http://www.ucmp.berkeley.edu/geology/tectonics.html>.

Kious, W. Jacquelyne, and Robert I. Tilling. *This Dynamic Earth: The Story of Plate Tectonics. U.S. Geological Survey* (Web site). <http://pubs.usgs.gov/publications/text/dynamic.html>.

Miller, Russell. *Continents in Collision.* Alexandria, VA: Time-Life Books, 1987.

Plate Tectonics (Web site). <http://www.platetectonics.com/>.

Plate Tectonics (Web site). <http://observe.ivv.nasa.gov/nasa/earth/tectonics/Tectonics1.html>.

Plate Tectonics, the Cause of Earthquakes (Web site). <http://www.seismo.unr.edu/ftp/pub/louie/class/100/plate-tectonics.html>.

Silverstein, Alvin, Virginia B. Silverstein, and Laura Silverstein Nunn. *Plate Tectonics.* Brookfield, CT: Twenty-First Century Books, 1998.

Wynn, Charles M., Arthur W. Wiggins, and Sidney Harris. *The Five Biggest Ideas in Science.* New York: John Wiley and Sons, 1997.

SEISMOLOGY

CONCEPT

Disturbances within Earth's interior, which is in a constant state of movement, result in the release of energy in packets known as seismic waves. An area of geophysics known as seismology is the study of these waves and their effects, which often can be devastating when experienced in the form of earthquakes. The latter do not only take lives and destroy buildings, but they also produce secondary effects, most often in the form of a tsunami, or tidal wave. Using seismographs and seismometers, seismologists study earthquakes and other seismic phenomena, including volcanoes and even explosions resulting from nuclear testing. They measure earthquakes according to their magnitude or energy as well as their intensity or human impact. Seismology also is used to study Earth's interior, about which it has revealed a great deal.

HOW IT WORKS

Stress and Strain in Earth's Interior

Modern earth scientists' studies in seismology, as in many other areas, are informed by plate tectonics, and to understand the causes of earthquakes and volcanoes, it is necessary to understand the basics of tectonics as well as plate tectonics theory. The latter subject is discussed in depth within a separate essay, which the reader is encouraged to consult for a more detailed explanation of concepts covered briefly here.

The term tectonism refers to deformation of the lithosphere, the upper layer of Earth's interior. Tectonics is the study of this deformation, which results from the release and redistribution of energy from Earth's core. The core is an extremely dense region, composed primarily of iron and another, lighter element (possibly sulfur), and is divided between a solid inner core with a radius of about 760 mi. (1,220 km) and a liquid outer core about 1,750 mi. (2,820 km) thick.

Earth's core possesses enormous energy, both gravitational and thermal. Gravitational energy is a result of the core's great mass (see Gravity and Geodesy for more about the role of mass in gravity), while thermal energy results from the radioactive decay of elements. In the context of radioactivity, decay does not mean "rot" rather, it refers to the release of high-energy particles. The release of these particles results in the generation of thermal energy, commonly referred to as heat. (See Energy and Earth for more about the scientific definition of heat as well as a discussion of geothermal energy.)

Differences in mass and temperature within the planet's interior, known as pressure gradients, result in the deformation of rocks in the lithosphere. The lithosphere includes the brittle upper portion of the mantle, a dense layer of rock approximately 1,429 mi. (2,300 km) thick, as well as the crust, which varies in depth from 3 mi. to 37 mi. (5–60 km). Deformation is the result of stress—that is, tension (stretching), compression, or shear. (The last of these stresses results from equal and opposite forces that do not act along the same line. To visualize shear, one need only imagine a thick hardbound book with its front cover pushed from the side so that the covers and pages are no longer perfectly aligned.)

A chasm along a fault scarp in San Bernardino County, California. *(© Ken M. Johns/Photo Researchers. Reproduced by permission.)*

Under the effects of these stresses, rocks experience strain, or a change in dimension as they bend, warp, slide, break, flow as though they were liquids, or melt. This strain, in turn, leads to a release of energy in the form of seismic waves. These waves may cause faults, or fractures, as well as folds, or bends in the rock structure, which manifest on the surface in the form of earthquakes, volcanoes, and other varieties of seismic activity. Seismology is the study of these waves as well as the movements and vibrations that produce them.

Continental Drift and Plate Tectonics

The theory of continental drift, discussed in Plate Tectonics, is based on the idea that the configuration of Earth's continents was once different than it is today. Integral to this theory is the accompanying idea that some of the individual land masses of today once were joined in other continental forms and that the land masses later moved to their present locations.

Continental drift theory was introduced in 1915 by the German geophysicist and meteorologist Alfred Wegener (1880–1930), but it failed to gain acceptance for half a century, in large part because it offered no explanation as to how the continents drifted. That explanation came in the 1960s with the development of plate tectonics, the name both of a theory and of a specialization of tectonics. As an area of study, plate tectonics deals with the large features of the lithosphere and the forces that shape them. As a theory, it explains the processes that have shaped Earth in terms of plates (large movable segments of the lithosphere) and their movement.

THE PLATES AND SEISMIC ACTIVITY. There are several major plates, some of which are listed in Plate Tectonics. That essay also discusses modern theories regarding the means by which continents broke apart many millions of years ago and then drifted back together, slamming into one another to form a number of notable features, such as the high mountains between the Indian subcontinent and the Eurasian landmass. Nor have the continents stopped moving; they continue to do so, though at a rate too slow to be noticed in a lifetime or even over the course of several generations. Based on its current rate of movement, in another 6,000 years—approximately the span of time since human civilization began—North America will have drifted about 600 ft. (183 m).

For the most part, the continents we know today are composed of single plates. For instance, South America sits on its own plate, which includes the southwestern quadrant of the Atlantic. But there are exceptions, an example being India itself, which is part of the Indo-Australian plate. Also notable is the Juan de Fuca Plate, a small portion of land attached to the North American continent and comprising the region from northern California to southern British Columbia.

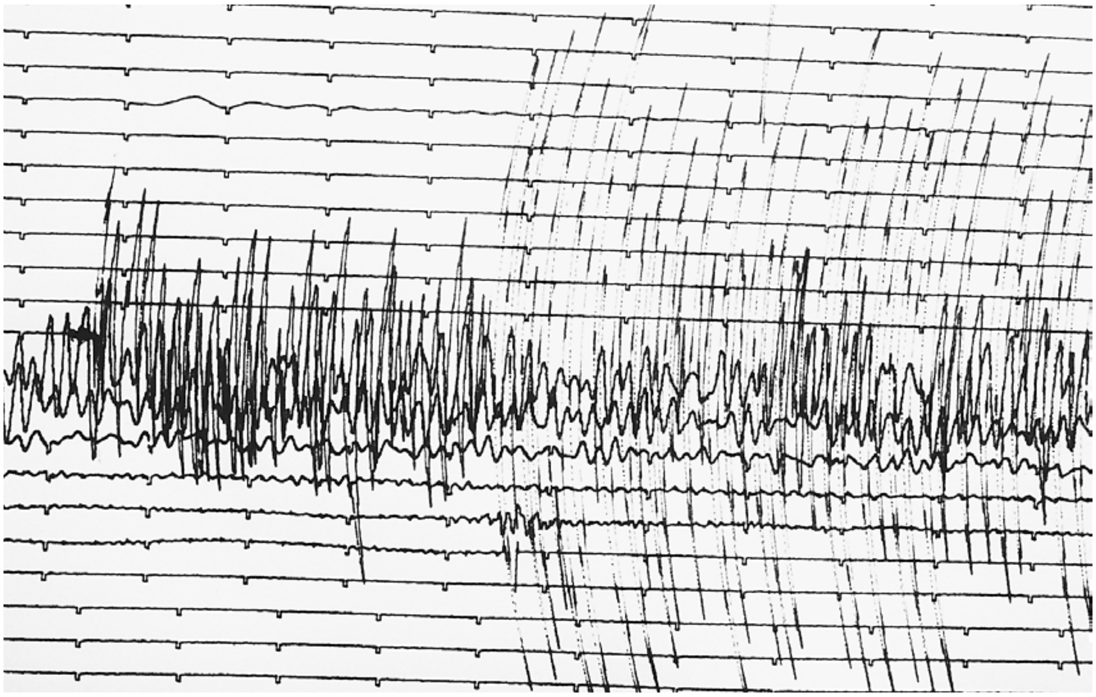
It so happens that this area is home to an unusual amount of volcanic activity. Southern California, where the North American and Pacific plates meet on the San Andreas fault, also is extremely prone to earthquakes, as is Japan, whose islands straddle the Philippine, Eurasian, and Pacific plates. Hawaii is another site of seismic activity in the form of volcanoes, but it does not lie at the nexus of any major plates. Instead, it is situated squarely atop the Pacific Plate, which is moving northward over a hot spot, a region of high volcanic activity. The hot spot remains more or less stationary, while the Pacific Plate moves across it; this explains why the volcanoes of northern Hawaii are generally dormant, whereas many volcanoes in the southern part of the island chain are still active.

Plates interact by moving toward each other (convergence), away from each other (divergence), or past each other (transform motion). When a continental plate converges with an oceanic plate (the differences between these types are discussed in Plate Tectonics), the much sturdier continental plate plows over the oceanic one. This is called subduction. The subducted plate undergoes partial melting, leading to the formation of volcanic chains, as in the nexus of the Juan de Fuca and Pacific or the South American and Nazca plates. The subduction of the Nazca Plate, which lies to the west of South America, helped form the Andes. Transform margins result in the formation of faults and earthquake zones, an example being the volatile San Andreas Fault.

Seismic Waves

The first scientific description of seismic waves was that of John William Strutt, Baron Rayleigh (1842–1919), who in 1885 characterized them as having aspects both of longitudinal and of transverse waves. These are, respectively, waves in which the movement of vibration is in the same direction as the wave itself and those in which the vibration or motion is perpendicular to the direction in which the wave is moving. (Ocean waves, for example, are longitudinal, whereas sound waves are transverse.)

Rayleigh waves later would be distinguished from Love waves, named after the English mathematician and geophysicist Augustus Edward Hough Love (1863–1940). The motion of Love waves is entirely horizontal, or longitudinal. Both are examples of surface waves, or seismic waves whose line of propagation is along the surface of a medium, such as the solid earth. These waves tend to be slower and more destructive than body waves, defined as waves whose line of propagation is through the body of a medium. Body waves include P-waves (primary waves), which are extremely fast moving and longitudinal, and S-waves (secondary waves), which move somewhat less fast and are transverse. The respective



A seismograph reading from the 1989 Loma Prieta, California, earthquake. *(© Russell D. Curtis/Photo Researchers. Reproduced by permission.)*

waves' rates of propagation through the solid earth are as follows:

- P-waves: about 4 mi. (6.4 km) per second
- S-waves: about 2 mi. (3.2 km) per second
- Rayleigh and Love waves: less than 2 mi. per second

REAL-LIFE APPLICATIONS

The Lisbon Quake and Its Effects

On November 1, 1755, the Portuguese capital of Lisbon became the site of one of the worst earthquakes in European history. The event had a number of aftereffects, natural and immediate as well as human and longer term. The results from nature were devastating; the earthquake caused a tsunami, or tidal wave, that flooded the Tagus River even as a fire, also caused by the earthquake, raged through the city.

Estimates of the deaths related directly or indirectly to the Lisbon quake range from 10,000 to as many as 60,000, making it the worst European earthquake since 1531. That earlier quake, incidentally, also had occurred in Lisbon—another example of the fact that certain areas are more prone to seismic activity. It so happens that Portugal lies near the boundary between the Eurasian and African plates.

As for the human response to the quake, it is best represented by the French writer Voltaire (1694–1778). Always a critic of religious faith, Voltaire saw in the incident evidence that called into question Christians' belief in a loving God. He made this case both in the philosophical poem *Le désastre de Lisbonne* (The disaster of Lisbon, 1756) and, more memorably, in the satirical novel *Candide* (1759).

MICHELL AND THE BIRTH OF SEISMOLOGY. Another, much less famous, thinker responded to the Lisbon earthquake in quite a different fashion. This was English geologist and astronomer John Michell (*ca.* 1724–1793), who studied the event and concluded that quakes are accompanied by shock waves. In an article published in 1760, he noted that earthquakes are found to occur near volcanoes and suggested that they are caused by pressure produced by water that boils from volcanic heat. He also indicated that one can calculate the center of an earthquake by making note of the time at which the motions are felt.

Today Michell is regarded as the father of seismology, a discipline that began to mature in the nineteenth century. The name itself was coined by the Irish engineer Robert Mallet (1810–1881), who in 1846 compiled the first modern catalogue of earthquakes. Eleven years after publishing the book, which listed all known quakes of any significance since 1606 B.C., Mallet conducted experiments with shock waves by exploding gunpowder and measuring the rate at which the waves travel through various types of material.

Detecting and Measuring Seismic Activity

As noted earlier, seismology is concerned with seismic waves, which generally are caused by movements within the solid earth. These waves also may be produced by man-made sources. Seismologic studies assist miners in knowing how much dynamite to use for a quarry blast so as to be effective without destroying the mine itself or the resources being sought. In addition, seismology can be used to reveal the location of such materials as coal and oil.

Thanks to seismometers (instruments for detecting seismic waves) and seismographs, which record information regarding those waves, seismologists are able to detect not only natural seismic activity but also the effects of underground nuclear testing. Underground testing is banned by international treaty, and if a "rogue nation" were to conduct such testing, it would come to the attention of the World-Wide Standardized Seismograph Network (WWSSN), which consists of 120 seismic stations in some 60 countries.

Most of the remainder of this essay is devoted to a single type of seismic phenomenon: earthquakes. As noted, they are far from the only effect of seismic activity; however, they are the most prevalent and well documented. A close second would be volcanoes, which are discussed in the essay Mountains.

EARLY SEISMOGRAPHIC INSTRUMENTS. In A.D. 132, the Chinese scientist Chang Heng (78–139) constructed what may have been the first seismographic instrument, which was designed to detect not only the presence of seismic activity but also the direction from which it came. His invention ultimately was discarded, however, and understanding of earthquakes progressed little for more than 1,600 years.

The first crude seismograph was invented in 1703 by the French physicist Jean de Hautefeuille (1647–1724), long before Michell formally established a connection between shock waves and earthquakes. Historians date the starting point of modern seismographic monitoring, however, to an 1880 invention by the English geologist John Milne (1850–1913). Milne's creation, the first precise seismograph, measured motion with a horizontal pendulum attached to a pen that recorded movement on a revolving drum. Milne used his device to record earthquakes from as far away as Japan and helped establish seismologic stations around the world. The first modern seismograph in the United States was installed at the University of California at Berkeley and proved its accuracy in recording the 1906 San Francisco quake, discussed later in this essay.

MAGNITUDE: THE RICHTER SCALE. An earthquake can be measured according to either its magnitude or its intensity. The first refers to the amount of energy released by the earthquake, and its best-known scale of measurement is the Richter scale. Developed in 1935 by the American geophysicist Charles Richter (1900–1985), the Richter scale is logarithmic rather than arithmetic, meaning that increases in value involve multiplication rather than addition.

The numbers on the Richter scale, from 1.0 to 10.0, should be thought of as exponents rather than integers. Each whole-number increase represents a tenfold increase in the amplitude (size from crest to trough) of the seismic wave. Therefore 2.0 is not twice as much as 1.0; it is 10 times as much. To go from 1.0 to 3.0 is an increase by a factor of 100, and to go from 1.0 to 4.0 indicates an increase by a factor of 1,000. The scales of magnitude thus become ever greater, and while a whole-number increase on the Richter scale indicates an increase of amplitude by a factor of 10, it represents an increase of energy by a factor of about 31.

INTENSITY: THE MERCALLI SCALE. The amplitude and energy measured by the Richter scale are objective and quantitative, whereas intensity is more subjective and qualitative. Intensity, an indication of the earthquake's effect on human beings and structures, is measured by the Mercalli scale, named after the Italian seismologist Giuseppe Mercalli (1850–1914). The 12 levels on the Mercalli scale range from I, which means that few people felt the quake, to XII, which indicates total damage. A few comparisons serve to illustrate the scales' relationship to each other.

A score of I on the Mercalli scale equates to a value between 1.0 and a 3.0 on the Richter scale and indicates a tremor felt only by a very few people under very specific circumstances. At 5.0 to 5.9 on the Richter scale (VI to VII on the Mercalli scale), everyone feels the earthquake, and many people are frightened, but only the most poorly built structures are damaged significantly. Above 7.0 on the Richter scale and VIII on the Mercalli scale, wooden and then masonry structures collapse, as do bridges, while railways are bent completely out of shape. In populated areas, as we shall see, the death toll can be enormous.

Famous Quakes

The great San Francisco earthquake, which struck on April 18, 1906, spawned a massive fire, and these events resulted in the deaths of some 700 people, including 270 inmates of a mental institution. Another 300,000 people were left homeless, and 490 city blocks were destroyed. Ultimately, the financial impact of the San Francisco quake proved to be one of the contributing factors in the March 13, 1907, stock market crash that played a key role in the panic of 1907.

At 5:04 P.M. on October 17, 1989, another quake struck San Francisco. It lasted just 15 seconds, long enough to kill some 90 people and cause $6 billion in property damage. Though it was the biggest quake since the 1906 tremor, it was much smaller: 7.19 on the Richter scale, or about one-fifth of the 7.7 measured for the 1906 quake. The 1989 Loma Prieta quake cost much more than the earlier tragedy, which had caused $500 million in damage, but, of course, half a billion dollars in 1906 was worth a great deal more than $6 billion 83 years later.

Neither earthquake, however, was the greatest in American history; in fact, the 1989 quake does not rank among the top 15, even for the continental United States. The eight worst earthquakes in U.S. history all occurred in one state: Alaska. Greatest of all was the March 27, 1964, quake at Prince William Sound, which registered a staggering 9.2 on the Richter scale and took 125 lives. Of that number, 110 were killed in a tsunami resulting from the quake.

The high incidence of earthquakes in Alaska is understandable enough, given the fact that its southern edge abuts a subduction zone and, along with the panhandle, sits astride the boundary between the North American and Pacific plates. Although this may not be much comfort to people in Alaska, it is fortunate that the most earthquake-prone state is also the most sparsely populated. Had the epicenter (the point on Earth's surface directly above the hypocenter, or focal point from which a quake originates) of the 1964 earthquake been in New York City, the death toll would have been closer to 125,000 than 125.

GREATEST QUAKES IN THE CONTINENTAL UNITED STATES. Similarly, it is fortunate that the greatest quakes to strike the continental United States outside California have been in low-population centers. Of the 15 worst earthquakes in U.S. history, only one was outside Alaska, California, or Hawaii. In fact, it was the site of both the worst and the fifth-worst earthquakes in the continental United States: New Madrid, Missouri, site of a 7.9 quake on February 7, 1812, and a 7.7 quake just two months earlier, on December 16, 1811.

New Madrid lies at the extreme southeastern tip of Missouri, near the Mississippi River and within a few hundred miles of several major cities: St. Louis, Missouri; Memphis and Nashville, Tennessee; and Louisville, Kentucky. Had the 1811 and 1812 quakes occurred today, they undoubtedly would have taken a vast human toll owing to the resulting floods. As it was, some lakes rose by as much as 15 ft. (4.6 m), streams changed direction, and the Mississippi and Ohio rivers flowed backward. Fortunately, however, they occurred at a time when the Missouri Territory—it was not even a state yet—and surrounding areas were sparsely populated. The combined death toll was in the single digits.

Of the top 15 earthquakes in the continental United States, all but the 1906 San Francisco quake (which ranks sixth) took place in areas with small populations. Ten were in California but generally in less populous areas or at times when there were fewer people there (e.g., no. 2: Fort Tejon, 1857; no. 3: Owens Valley, 1872; and no. 4: Imperial Valley, 1892). Other than the two New Madrid quakes, the remainder took place in Nevada (no. 12: Dixie Valley, 1954), Montana (no. 13: Hebgen Lake, 1959), and Idaho (no. 14:

Earthquake damage in California. *(© David Weintraub/Photo Researchers. Reproduced by permission.)*

Borah Peak, 1983). As of late 2001, the Idaho quake was the second most recent, after no. 9, at Landers, California, in 1992. (The 1994 Northridge quake, in the Los Angeles area, ranked 6.7 on the Richter scale, well below the 7.3 registered by no. 15, west of Eureka, California, in 1922.)

THE WORLD'S MOST DESTRUCTIVE QUAKES. None of these U.S. quakes, however, compares with the July 27, 1976, earthquake in T'ang-shan, China. The worst earthquake in modern history, it shattered some 20 sq. mi. (32 km sq.) near the capital city of Beijing and killed about 242,000 people while injuring an estimated 600,000 more. There are several interesting aspects to this quake, aside from its sheer scale.

One is sociological, involving the human response to the quake. As in Portugal in 1755, people saw events in a cosmic light; in this case, though, they did not interpret the quake as evidence of divine unconcern but quite the opposite. Mao Tse-tung (1893–1976), by far the most influential Chinese leader of modern times, had just died, and the Chinese saw the natural disaster as fitting into a larger historical pattern. In the traditional Chinese view, earthquakes, floods, and other signs from the gods attend the change of dynasties.

Also interesting is the fact that the T'ang-shan quake was merely the most destructive in a worldwide series of quakes that took place between February and November 1976. In the course of these events, 23,000 people died in Guatemala after a February 4 quake; 3,000 people were reported dead, and 3,000 more were missing in Indonesia, as a result of a series of quakes and landslides on June 26 (later, the U.S. Federal Emergency Management Agency, or FEMA, placed the number of dead from the Indonesia quake at just 443); as many as 8,000 people died in an earthquake and tsunami that hit the southern Philippines on August 16; and 4,000 more perished in a November 24 quake in eastern Turkey.

Similarly, a few months before the 1755 Lisbon earthquake, a quake hit northern Iran. This is an aspect of seismology that cannot be explained readily by plate tectonics: Iran and Portugal are not on the same plate margins; in fact, northern Iran is not on a plate margin at all. Likewise, the areas hit in the 1976 quakes were not on the same plate margins, and T'ang-shan (unlike the other places affected) is not on a major plate margin at all. Nor is Shansi in north-central China, site of history's most destructive earthquake on January 24, 1556, which killed more than 830,000 people.

Note that the 1556 and 1976 Chinese quakes were the worst, respectively, of all history and of modern times—but *worst* in terms of intensity, not magnitude. One might say that they were the most destructive but not the worst in pure terms. The 1976 quake is not even on the list of the 10 worst earthquakes—those of the greatest magnitude—in the twentieth century. Whereas the T'ang-shan quake registered 8.0, a quake in Chile on May 22, 1960, had a magnitude of 9.5, or about 50 times greater, yet the death toll was much smaller—2,000 people killed. Three thousand more were injured in the Chilean quake, and two million were rendered homeless. The last statistic perhaps best signifies the magnitude of the 1960 quake, which caused tsunamis that brought death and destruction as far away as Hawaii, Japan, the Philippines, and the west coast of the United States.

Learning from Seismology

As noted, plate tectonics does not explain every earthquake, but it does explain most, probably about 90%. Not that it is much help in predicting earthquakes, because the processes of plate tectonics take place on an entirely different time scale than the ones to which humans are accustomed. These processes happen over millions of years, so it is hard to say, for any particular year, just what will happen to a particular plate.

Plate tectonics, then, tells us only areas of likelihood for earthquakes—specifically, plate boundaries of the types discussed near the end of Plate Tectonics. And even though the processes that create the conditions for an earthquake are extremely slow, usually the discernible indications that an earthquake is coming appear only seconds before the quake itself. Thus, as sophisticated as modern seismometers are, they generally do not provide enough advance notice of earthquakes to offer any lifesaving value.

There are not just a few earthquakes each year but many thousands of tremors, most of them too small to register. Sometimes these tremors may be foreshocks, or indicators that a quake is coming to a particular area. In addition, studies of other phenomena, from tidal behavior to that of animals (probably a result of some creatures' extremely acute hearing), may offer suggestions as to the locations of future quakes.

EARTH'S CORE AND THE MOHO. Seismology is useful for learning about more than just earthquakes or volcanoes. During the early years of the twentieth century, the Irish geologist Richard Dixon Oldham (1858–1936) studied data from a number of recent earthquakes and noticed a difference in the behavior of compression waves and shear waves. (These terms merely express the differ-

KEY TERMS

AMPLITUDE: The maximum displacement of a vibrating material, or the "size" of a wave from crest to trough.

BODY WAVES: Waves whose line of propagation is through the body of a medium. These include P-waves (primary waves), which move extremely fast and are longitudinal, and S-waves (secondary waves), which are move somewhat less fast and are transverse. Compare with *surface waves.*

COMPRESSION: A form of stress produced by the action of equal and opposite forces, the effect of which is to reduce the length of a material. Compression is a form of pressure.

CONTINENTAL DRIFT: The theory that the configuration of Earth's continents was once different than it is today, that some of the individual landmasses of today once were joined in other continental forms, and that these landmasses later separated and moved to their present locations.

CONVERGENCE: A tectonic process whereby plates move toward each other. Usually associated with subduction, convergence typically occurs in the ocean, creating an oceanic trench. It is one of the three ways, along with divergence and transform motion, that plates interact.

CORE: The center of Earth, an area constituting about 16% of the planet's volume and 32% of its mass. Made primarily of iron and another, lighter element (possibly sulfur), it is divided between a solid inner core with a radius of about 760 mi. (1,220 km) and a liquid outer core about 1,750 mi. (2,820 km) thick.

CRUST: The uppermost division of the solid earth, representing less than 1% of its volume and varying in depth from 3–37 mi. (5–60 km). Below the crust is the mantle.

DIVERGENCE: A tectonic process whereby plates move away from each other. Divergence results in the separation of plates and most often is associated either with seafloor spreading or the formation of rift valleys. It is one of the three ways, along with convergence and transform motion, that plates interact.

ELASTICITY: The response of solids to stress.

EPICENTER: The point on Earth's surface directly above the hypocenter, or the focal point from which an earthquake originates.

FAULT: An area of fracturing, as a result of stress, between rocks.

FOLD: An area of rock that has been bent by stress.

GEOPHYSICS: A branch of the earth sciences that combines aspects of geology and physics. Geophysics addresses the planet's physical processes as well as its gravitational, magnetic, and electric properties and the means by which energy is transmitted through its interior.

HEAT: Internal thermal energy that flows from one body of matter to another.

HOT SPOT: A region of high volcanic activity.

KEY TERMS CONTINUED

INTENSITY: Where earthquakes are concerned, intensity refers to the amount of damage to humans and buildings. Subjective and qualitative (as opposed to magnitude, which is objective and quantitative), intensity is measured by the Mercalli scale.

KINETIC ENERGY: The energy that an object possesses by virtue of its motion.

LITHOSPHERE: The upper layer of Earth's interior, including the crust and the brittle portion at the top of the mantle.

LONGITUDINAL WAVE: A wave in which the movement of vibration is in the same direction as the wave itself. This is contrasted with a *transverse wave.*

LOVE WAVES: See *surface waves.*

MAGNITUDE: Where earthquakes are concerned, magnitude refers to the amount of energy released by the quake as well as the amplitude of the seismic waves. Objective and quantitative (as opposed to intensity, which is subjective and qualitative), magnitude is measured by the Richter scale.

MANTLE: The thick, dense layer of rock, approximately 1,429 mi. (2,300 km) thick, between Earth's crust and its core. In reference to the other terrestrial planets, *mantle* simply means the area of dense rock between the crust and core.

MERCALLI SCALE: See *intensity.*

PLATE MARGINS: Boundaries between plates.

PLATE TECTONICS: The name both of a theory and of a specialization of tectonics. As an area of study, plate tectonics deals with the large features of the lithosphere and the forces that shape them. As a theory, it explains the processes that have shaped Earth in terms of plates and their movement. Plate tectonics theory brings together aspects of continental drift, seafloor spreading, seismic and volcanic activity, and the structures of Earth's crust to provide a unifying model of Earth's evolution. It is one of the dominant concepts in the modern earth sciences.

PLATES: Large, movable segments of the lithosphere.

PROPAGATION: The act or state of traveling from one place to another.

P-WAVES: See *body waves.*

RADIOACTIVITY: A term describing a phenomenon whereby certain materials are subject to a form of decay brought about by the emission of high-energy particles or radiation. Forms of particles or energy include alpha particles (positively charged helium nuclei); beta particles (either electrons or subatomic particles called *positrons*); or gamma rays, which occupy the highest energy level in the electromagnetic spectrum.

RAYLEIGH WAVES: See *surface waves.*

RICHTER SCALE: See *magnitude.*

SEISMIC WAVE: A packet of energy resulting from the disturbance that accompanies a strain on rocks in the lithosphere.

SEISMOGRAPH: An instrument designed to record information regarding seismic waves.

KEY TERMS CONTINUED

SEISMOLOGY: The study of seismic waves as well as the movements and vibrations that produce them.

SEISMOMETER: An instrument for detecting seismic waves.

SHEAR: A form of stress resulting from equal and opposite forces that do not act along the same line. If a thick, hardbound book is lying flat and one pushes the front cover from the side so that the covers and pages no longer constitute parallel planes, this is an example of shear.

STRAIN: The ratio between the change in dimension experienced by an object that has been subjected to stress and the original dimensions of the object.

STRESS: In general terms, any attempt to deform a solid. Types of stress include tension, compression, and shear.

SUBDUCTION: A tectonic process that results when plates converge, and one plate forces the other down into Earth's mantle. As a result, the subducted plate eventually undergoes partial melting.

SURFACE WAVES: Seismic waves whose line of propagation is along the surface of a medium such as the solid earth. These waves tend to be slower and more destructive than body waves. Examples include Rayleigh waves (waves with both transverse and longitudinal characteristics) and Love waves (purely longitudinal). Compare with *body waves.*

S-WAVES: See *body waves.*

TECTONICS: The study of tectonism, including its causes and effects, most notably mountain building.

TECTONISM: The deformation of the lithosphere.

TENSION: A form of stress produced by a force that acts to stretch a material.

THERMAL ENERGY: Heat energy, a form of kinetic energy produced by the motion of atomic or molecular particles in relation to one another. The greater the relative motion of these particles, the greater the thermal energy.

TRANSFORM MOTION: A tectonic process whereby plates slide past each other. It is one of the three ways, along with convergence and divergence, that plates interact.

TRANSVERSE WAVE: A wave in which the vibration or motion is perpendicular to the direction in which the wave is moving. Compare with *longitudinal wave.*

TSUNAMI: A tidal wave produced by an earthquake or volcanic eruption. The term comes from the Japanese words for "harbor" and "wave."

ences in stress produced by seismic waves.) As it turns out, shear waves are deflected as they pass through the center of Earth. Since liquid cannot experience shear, this finding told him that the planet's core must be made of molten material.

Oldham's findings, published in 1906—the same year as the great San Francisco quake—made him a pioneer in the application of seismology to the study of Earth's interior. Three years later, studies of earthquake waves by the Croatian geologist Andrija Mohorovicic (1857–1936) revealed still more about the interior of the planet. Based on his analysis of wave speeds and arrival times, Mohorovicic was able to calculate the depth at which the crust becomes the mantle. This change is abrupt rather than gradual, and the boundary on which it occurs is today known as the Mohorovicic discontinuity, or simply the Moho.

WHERE TO LEARN MORE

Abbott, Patrick. *Natural Disasters.* Dubuque, IA: William C. Brown Publishers, 1996.

National Earthquake Information Center—World Data Center for Seismology, Denver (Web site). <http://neic.usgs.gov/>.

Prager, Ellen J., Kate Hutton, Costas Synolakis, et al. *Furious Earth: The Science and Nature of Earthquakes, Volcanoes, and Tsunamis.* New York: McGraw-Hill, 2000.

Seismology Info Page—Netherlands (Web site). <http://home.wish.net/~riknl/>.

Seismology Research Centre—Australia (Web site). <http://www.seis.com.au/>.

Sigurdsson, Haraldur. *Encyclopedia of Volcanoes.* San Diego: Academic Press, 2000.

Sleh, Keery E., and Simon LeVay. *The Earth in Turmoil: Earthquakes, Volcanoes, and Their Impact on Humankind.* New York: W. H. Freeman, 1998.

University of Alaska Fairbanks Seismology (Web site). <http://www.aeic.alaska.edu/>.

University of Washington Seismology and Earthquake Information (Web site). <http://www.geophys.washington.edu/SEIS/>.

"Earthquake Hazards Program," USGS (United States Geological Survey) (Web site). <http://earthquake.usgs.gov/>.

Wade, Nicholas. *The Science Times Book of Natural Disasters.* New York: Lyons Press, 2000.

GEOMORPHOLOGY

GEOMORPHOLOGY

CONCEPT

The surface of Earth is covered with various landforms, a number of which are discussed in various entries throughout this book. This essay is devoted to the study of landforms themselves, a subdiscipline of the geologic sciences known as geomorphology. The latter, as it has evolved since the end of the nineteenth century, has become an interdisciplinary study that draws on areas as diverse as plate tectonics, ecology, and meteorology. Geomorphology is concerned with the shaping of landforms, through such processes as subsidence and uplift, and with the classification and study of such landforms as mountains, volcanoes, and islands.

HOW IT WORKS

An Evolving Area of Study

Geomorphology is an area of geology concerned with the study of landforms, with the forces and processes that have shaped them, and with the description and classification of various physical features on Earth. The term, which comes from the Greek words *geo,* or "Earth," and *morph,* meaning "form," was coined in 1893 by the American geologist William Morris Davis (1850–1934), who is considered the father of geomorphology.

During Davis's time, geomorphology was concerned primarily with classifying different structures on Earth's surface, examples of which include mountains and islands, discussed later in this essay. This view of geomorphology as an essentially descriptive, past-oriented area of study closely aligned with historical geology prevailed throughout the late nineteenth and early twentieth centuries.

By the mid-twentieth century, however, the concept of geomorphology inherited from Davis had fallen into disfavor, to be replaced by a paradigm, or model, oriented toward physical rather than historical geology. (These two principal branches of geology are concerned, in the first instance, with Earth's past and the processes that shaped it and, in the second instance, with Earth's current physical features and the processes that continue to shape it.)

RETHINKING GEOMORPHOLOGY. As reconceived in the 1950s and thereafter, geomorphology became an increasingly exact science. As has been typical of many sciences in their infancy, early geomorphology focused on description rather than prediction and tended to approach its subject matter in a qualitative fashion. The term qualitative suggests a comparison between qualities that are not defined precisely, such as "fast" and "slow" or "warm" and "cold." On the other hand, a quantitative approach, as has been implemented for geomorphology from the mid–twentieth century onward, centers on a comparison between precise quantities—for instance, 10 lb. (4.5 kg) versus 100 lb. (45 kg) or 50 MPH (80.5 km/h) versus 120 MPH (193 km/h).

As part of its shift in focus, geomorphology began to treat Earth's physical features as systems made up of complex and ongoing interactions. This view fell into line with a general emphasis on the systems concept in the study of Earth. (See Earth Systems for more about the systems concept.) As geomorphology evolved, it became more interdisciplinary, as we shall see. This, too,

MOUNT MACHHAPUCHHARE IN THE HIMALAYAS. THE HIMALAYAS WERE FORMED THROUGH THE PROCESS OF UPLIFT, OR RISING OF EARTH'S SURFACE. *(© George Turner/Photo Researchers. Reproduced by permission.)*

was part of an overall trend in the earth sciences toward an approach that viewed subjects in broad, cross-disciplinary terms as opposed to a narrow focus on specific areas of study.

LANDFORMS AND PROCESSES

Two concerns are foremost within the realm of geomorphology, and these concerns reflect the stages of its history. First, in line with Davis's original conception of geomorphology as an area of science devoted to classifying and describing natural features, there is its concern with topography. The latter may be defined as the configuration of Earth's surface, including its relief (elevation and other inequalities) as well as the position of physical features.

These physical features are called landforms, examples of which include mountains, plateaus, and valleys. Geomorphology always has involved classification, and early scientists working in this subdiscipline addressed the classification of landforms. Other systems of classification, however, are not so concerned with cataloging topographical features themselves as with differentiating the processes that shaped them. This brings us to the other area of interest in geomorphology: the study of how landforms came into being.

SHAPING THE EARTH. Among the processes that drive the shaping of landforms is plate tectonics, or the shifting of large, movable segments of lithosphere (the crust and upper layer of Earth's mantle). Plate tectonics is discussed in detail within its own essay and more briefly in other areas throughout this book, as befits its status as one of the key areas of study in the earth sciences.

Other processes also shape landforms. Included among these processes are weathering, the breakdown of rocks and minerals at or near the surface of Earth due to physical or chemical processes; erosion, the movement of soil and rock due to forces produced by water, wind, glaciers, gravity, and other influences; and mass wasting or mass movement, the transfer of earth material, by processes that include flow, slide, fall, and creep, down slopes. Also of interest are fluvial and eolian processes (those that result from water flow and wind, respectively) as well as others related to glaciers and coastal formations.

Human activity also can play a significant role in shaping Earth. This effect may be direct, as when the construction of cities, the building of dams, or the excavation of mines alters the landscape. On the other hand, it can be indirect. In the latter instance, human activity in the biosphere exerts an impact, as when the clearing of forest land or the misuse of crop land results in the formation of a dust bowl.

Interdisciplinary Studies

As noted earlier, geomorphology is characteristic of the earth sciences as a whole in its emphasis on an interdisciplinary approach. As is true of earth scientists in general, those studying landforms and the processes that shape them do not work simply in one specialty. Among the areas of interest in geomorphology are, for example, deep-sea geomorphology, which draws on oceanography, and planetary geomorphology, the study of landscapes on other planets.

When studying coastal geomorphology, a geologist may draw on realms as diverse as fluid mechanics (an area of physics that studies the behavior of gases and liquids at rest and in motion) and sedimentology. The investigation of such processes as erosion and mass wasting calls on knowledge in the atmospheric sciences as well as the physics and chemistry of soil. It is almost inevitable that a geomorphologic researcher will draw on geophysics as well as on such subspecialties as volcanology. These studies may go beyond the "hard sciences," bringing in such social sciences as geography.

REAL-LIFE APPLICATIONS

Subsidence

Subsidence refers to the process of subsiding (settling or descending), on the part of either an air column or the solid earth, or, in the case of solid earth, to the resulting formation or depression. Subsidence in the atmosphere is discussed briefly in the entry Convection. Subsidence that occurs in the solid earth, known as geologic subsidence, is the settling or sinking by a body of rock or sediment. (The latter can be defined as material deposited at or near Earth's surface from a number of sources, most notably preexisting rock.)

As noted earlier, many geomorphologic processes can be caused either by nature or by human beings. An example of natural subsidence takes place in the aftermath of an earthquake, during which large areas of solid earth may simply drop by several feet. Another example can be observed at the top of a volcano some time after it has erupted, when it has expelled much of its material (i.e., magma) and, as a result, has collapsed.

Natural subsidence also may result from cave formation in places where underground water has worn away limestone. If the water erodes too much limestone, the ceiling of the cave will subside, usually forming a sinkhole at the surface. The sinkhole may fill with water, making a lake; the formation of such sinkholes in many spots throughout an area (whether the sinkholes become lakes or not), is known as karst topography.

In places where the bedrock is limestone—particularly in the sedimentary basins of rivers—karst topography is likely to develop. The United States contains the most extensive karst region in the world, including the Mammoth cave system in Kentucky. Karst topography is very pronounced in the hills of southern China, and karst landscapes have been a prominent feature of Chinese art for centuries. Other extensive karst regions can be found in southern France, Central America, Turkey, Ireland, and England.

MAN-MADE SUBSIDENCE. Man-made subsidence often ensues from the removal of groundwater or fossil fuels, such as petroleum or coal. Groundwater removal can be perfectly safe, assuming the area experiences sufficient rainfall to replace, or recharge, the lost water. If recharging does not occur in the necessary proportions, however, the result will be the eventual collapse of the aquifer, a layer of rock that holds groundwater.

In so-called room-and-pillar coal mining, pillars, or vertical columns, of coal are left standing, while the areas around them are extracted. This method maintains the ceiling of the "room" that has been mined of its coal. After the mine is abandoned, however, the pillar eventually may experience so much stress that it breaks, leading to the collapse of the mined room. As when the ceiling of a cave collapses, the subsidence of a coal mine leaves a visible depression above ground.

UPLIFT

As its name implies, uplift describes a process and results opposite to those of subsidence. In uplift the surface of Earth rises, owing either to a decrease in downward force or to an increase in upward force. One of the most prominent examples of uplift is seen when plates collide, as when India careened into the southern edge of the Eurasian landmass some 55 million years ago. The result has been a string of mountain ranges, including the Himalayas, Karakoram Range, and Hindu Kush, that contain most of the world's tallest peaks.

Plates move at exceedingly slow speeds, but their mass is enormous. This means that their inertia (the tendency of a moving object to keep moving unless acted upon by an outside force) is likewise gargantuan in scale. Therefore, when plates collide, though they are moving at a rate equal to only a few inches a year, they will keep pushing into each other like two automobiles crumpling in a head-on collision. Whereas a car crash is over in a matter of seconds, however, the crumpling of continental masses takes place over hundreds of thousands of years.

When sea floor collides with sea floor, one of the plates likely will be pushed under by the other one, and, likewise, when sea floor collides with continental crust, the latter will push the sea floor under. (See Plate Tectonics for more about oceanic-oceanic and continental-oceanic collisions.) This results in the formation of volcanic mountains, such as the Andes of South America or the Cascades of the Pacific Northwest, or volcanic islands, such as those of Japan, Indonesia, or Alaska's Aleutian chain.

ISOSTATIC COMPENSATION. In many other instances, collision, compression, and extension cause uplift. On the other hand, as noted, uplift may result from the removal of a weight. This occurs at the end of an ice age, when glaciers as thick as 1.9 mi. (3 km) melt, gradually removing a vast weight pressing down on the surface below.

This movement leads to what is called isostatic compensation, or isostatic rebound, as the crust pushes upward like a seat cushion rising after a person is longer sitting on it. Scandinavia is still experiencing uplift at a rate of about 0.5 in. (1 cm) per year as the after-effect of glacial melting from the last ice age. The latter ended some 10,000 years ago, but in geologic terms this is equivalent to a few minutes' time on the human scale.

ISLANDS

Geomorphology, as noted earlier, is concerned with landforms, such as mountains and volcanoes as well as larger ones, including islands and even continents. Islands present a particularly interesting area of geomorphologic study. In general, islands have certain specific characteristics in terms of their land structure and can be analyzed from the standpoint of the geosphere, but particular islands also have unique ecosystems, requiring an interdisciplinary study that draws on botany, zoology, and other subjects.

In addition, there is something about an island that has always appealed to the human imagination, as evidenced by the many myths, legends, and stories about islands. Some examples include Homer's *Odyssey,* in which the hero Odysseus visits various islands in his long wanderings; Thomas More's *Utopia,* describing an idealized island republic; *Robinson Crusoe,* by Daniel Defoe, in which the eponymous hero lives for many years on an island with no companion but the trusty native Friday; *Treasure Island,* by Robert Louis Stevenson, in which the island is the focus of a treasure hunt; and Mark Twain's *Adventures of Huckleberry Finn,* depicting Jackson Island in the Mississippi River, to which Huckleberry Finn flees to escape "civilization."

One of the favorite subjects of cartoonists is that of a castaway stranded on a desert island, a mound of sand with no more than a single tree.

Movies, too, have long portrayed scenarios, from the idyllic to the brutal, that take place on islands, particularly deserted ones, a notable example being *Cast Away* (2000). A famous line by the English poet John Donne (1572–1631) warns that "no man is an island," implying that many wish they could enjoy the independence suggested by the concept of an island. Within the Earth system, however, nothing is fully independent, and, as we shall see, this is certainly the case where islands are concerned.

THE ISLANDS OF EARTH. Earth has literally tens of thousands of islands. Just two archipelagos (island chains), those that make up the Philippines and Indonesia, include thousands of islands each. While there are just a few dozen notable islands on Earth, many more dot the planet's seas and oceans. The largest are these:

- Greenland (Danish, northern Atlantic): 839,999 sq. mi.(2,175,597 sq km)
- New Guinea (divided between Indonesia and Papua New Guinea, western Pacific): 316,615 sq. mi. (820,033 sq km)
- Borneo (divided between Indonesia and Malaysia, western Pacific): 286,914 sq. mi. (743,107 sq km)
- Madagascar (Malagasy Republic, western Indian Ocean): 226,657 sq. mi. (587,042 sq km)
- Baffin (Canadian, northern Atlantic): 183,810 sq. mi. (476,068 sq km)
- Sumatra (Indonesian, northeastern Indian Ocean): 182,859 sq. mi. (473,605 sq km)

The list could go on and on, but it stops at Sumatra because the next-largest island, Honshu (part of Japan), is less than half as large, at 88,925 sq. mi. (230,316 sq km). Clearly, not all islands are created equal, and though some are heavily populated or enjoy the status of independent nations (e.g., Great Britain at number eight or Cuba at number 15), they are not necessarily the largest. On the other hand, some of the largest are among the most sparsely populated.

Of the 32 largest islands in the world, more than a third are in the icy northern Atlantic and Arctic, with populations that are small or practically nonexistent. Greenland's population, for instance, was just over 59,000 in 1998, while that of Baffin Island was about 13,200. On both islands, then, each person has about 14 frozen sq. mi. (22 sq km) to himself or herself, making them among the most sparsely populated places on Earth.

A tiny island in the Truk lagoon, Micronesia. *(© Stuart Westmorland/Photo Researchers. Reproduced by permission.)*

CONTINENTS, OCEANS, AND ISLANDS. Australia, of course, is not an island but a continent, a difference that is not related directly to size. If Australia *were* an island, it would be by far the largest. Australia is regarded as a continent, however, because it is one of the principal landmasses of the Indo-Australian plate, which is among a handful of major continental plates on Earth. Whereas continents are more or less permanent (though they have experienced considerable rearrangement over the eons), islands come and go, seldom lasting more than 10 million years. Erosion or rising sea levels remove islands, while volcanic explosions can create new ones, as when an eruption off the coast of Iceland resulted in the formation of an island, Surtsey, in 1963.

Islands are of two types, continental and oceanic. Continental islands are part of continental shelves (the submerged, sloping ledges of continents) and may be formed in one of two ways. Rising ocean waters either cover a coastal region, leaving only the tallest mountains exposed as islands or cut off part of a peninsula,

KEY TERMS

BIOSPHERE: A combination of all living things on Earth—plants, animals, birds, marine life, insects, viruses, single-cell organisms, and so on—as well as all formerly living things that have not yet decomposed.

CONVERGENCE: A tectonic process whereby plates move toward each other. Usually associated with subduction, convergence typically occurs in the ocean, creating an oceanic trench. It is one of the three ways, along with divergence and transform motion, in which plates interact.

CRUST: The uppermost division of the solid earth, representing less than 1% of its volume and varying in depth from 3 to 37 mi. (5 to 60 km). Below the crust is the mantle.

EROSION: The movement of soil and rock due to forces produced by water, wind, glaciers, gravity, and other influences.

GEOLOGY: The study of the solid earth, in particular its rocks, minerals, fossils, and land formations.

GEOMORPHOLOGY: An area of geology concerned with the study of landforms, with the forces and processes that have shaped them, and with the description and classification of various physical features on Earth.

GEOPHYSICS: A branch of the earth sciences that combines aspects of geology and physics. Geophysics addresses the planet's physical processes as well as its gravitational, magnetic, and electric properties and the means by which energy is transmitted through its interior.

GEOSPHERE: The upper part of Earth's continental crust, or that portion of the solid earth on which human beings live and which provides them with most of their food and natural resources.

HISTORICAL GEOLOGY: The study of Earth's physical history. Historical geology is one of two principal branches of geology, the other being physical geology.

LANDFORM: A notable topographical feature, such as a mountain, plateau, or valley.

LITHOSPHERE: The upper layer of Earth's interior, including the crust and the brittle portion at the top of the mantle.

MASS WASTING: The transfer of earth material, by processes that include flow, slide, fall, and creep, down slopes. Also known as mass movement.

which then becomes an island. Most of Earth's significant islands are continental and are easily spotted as such, because they lie at close proximity to continental landmasses. Many other continental islands are very small, however; examples include the barrier islands that line the East Coast of the United States. Formed from mainland sand brought to the coast by rivers, these are technically not continental islands, but they more clearly fit into that category than into the grouping of oceanic islands.

Oceanic islands, of which the Hawaiian-Emperor island chain and the Aleutians off the Alaskan coast are examples, form as a result of volcanic activity on the ocean floor. In most cases, there is a region of high volcanic activity,

KEY TERMS CONTINUED

PHYSICAL GEOLOGY: The study of the material components of Earth and of the forces that have shaped the planet. Physical geology is one of two principal branches of geology, the other being historical geology.

PLATE TECTONICS: The name both of a theory and of a specialization of tectonics. As an area of study, plate tectonics deals with the large features of the lithosphere and the forces that shape them. As a theory, it explains the processes that have shaped Earth in terms of plates and their movement.

PLATES: Large, movable segments of the lithosphere.

QUALITATIVE: Involving a comparison between qualities that are not defined precisely, such as "fast" and "slow" or "warm" and "cold."

QUANTITATIVE: Involving a comparison between precise quantities—for instance, 10 lb. versus 100 lb. or 50 mi. per hour versus 120 mi. per hour.

RELIEF: Elevation and other inequalities on a land surface.

SEDIMENT: Material deposited at or near Earth's surface from a number of sources, most notably preexisting rock.

SEDIMENTOLOGY: The study and interpretation of sediments, including sedimentary processes and formations.

SUBSIDENCE: A term that refers either to the process of subsiding (settling or descending), on the part of either air or solid earth or, in the case of solid earth, to the resulting formation. Subsidence thus is defined variously as the downward movement of air, the sinking of ground, or a depression in Earth's crust.

SYSTEM: Any set of interactions that can be set apart mentally from the rest of the universe for the purposes of study, observation, and measurement.

TECTONICS: The study of tectonism, including its causes and effects, most notably mountain building.

TECTONISM: The deformation of the lithosphere.

TOPOGRAPHY: The configuration of Earth's surface, including its relief, as well as the position of physical features.

UPLIFT: A process whereby the surface of Earth rises, due to either a decrease in downward force or an increase in upward force.

WEATHERING: The breakdown of rocks and minerals at or near the surface of Earth due to physical or chemical processes.

called a hot spot, beneath the plates, which move across the hot spot. This is the situation in Hawaii, and it explains why the volcanoes on the southern islands are still active while those to the north are not: the islands themselves are moving north across the hot spot. If two plates converge and one subducts (see Plate Tectonics for an explanation of this process), a deep trench with a parallel chain of volcanic islands may develop. Exemplified by the Aleutians, these chains are called island arcs.

ISLAND ECOSYSTEMS. The ecosystem, or community of all living organisms, on islands can be unique owing to their separation from continents. The number of life-forms on an island is relatively small and can encompass some unusual circumstances compared with the larger ecosystems of continents. Ireland, for instance, has no native snakes, a fact "explained" by the legend that Saint Patrick drove them away. Hawaii and Iceland are also blessedly free of serpents.

Oceanic islands, of course, tend to have more unique ecosystems than do continental islands. The number of land-based animal lifeforms is necessarily small, whereas the varieties of birds, flying insects, and surrounding marine life will be greater owing to those creatures' mobility across water. Vegetation is relatively varied, given the fact that winds, water currents, and birds may carry seeds.

Nonetheless, ecosystems of islands tend to be fairly delicate and can be upset by the human introduction of new predators (e.g., dogs) or new creatures to consume plant life (e.g., sheep). These changes sometimes can have disastrous effects on the overall balance of life on islands. Overgrazing may even open up the possibility of erosion, which has the potential of bringing an end to an island's life.

WHERE TO LEARN MORE

Color Landform Atlas of the United States (Web site). <http://fermi.jhuapl.edu/states/states.html>.

Erickson, Jon. *Rock Formations and Unusual Geologic Structures: Exploring the Earth's Surface.* New York: Facts on File, 1993.

Erickson, Jon. *Making the Earth: Geologic Forces That Shape Our Planet.* New York: Facts on File, 2000.

Gerrard, John. *Rocks and Landforms.* Boston: Unwin Hyman, 1988.

Image Gallery of Landforms (Web site). <http://www.athena.ivv.nasa.gov/curric/land/landform/landform.html>.

Selby, Michael John. *Earth's Changing Surface: An Introduction to Geomorphology.* New York: Clarendon Press, 1985.

Tinkler, K. J. *A Short History of Geomorphology.* Totowa, NJ: Barnes and Noble, 1985.

The Virtual Geomorphology (Web site). <http://main.amu.edu.pl/~sgp/gw/gw.htm>.

Wells, Lisa. "Images Illustrating Principles of Geomorphology" (Web site). <http://geoimages.berkeley.edu/GeoImages/Wells/wells.html>.

Zoehfeld, Kathleen Weidner, and James Graham Hale. *How Mountains Are Made.* New York: HarperCollins, 1995.

MOUNTAINS

CONCEPT

Among the most striking of geologic features are mountains, created by several types of tectonic forces, including collisions between continental masses. Mountains have long had an impact on the human psyche, for instance by virtue of their association with the divine in the Greek myths, the Bible, and other religious or cultural traditions. One does not need to be a geologist to know what a mountain is; indeed there is no precise definition of mountain, though in most cases the distinction between a mountain and a hill is fairly obvious. On the other hand, the defining characteristics of a volcano are more apparent. Created by violent tectonic forces, a volcano usually is considered a mountain, and almost certainly is one after it erupts, pouring out molten rock and other substances from deep in the earth.

HOW IT WORKS

Plate Tectonics

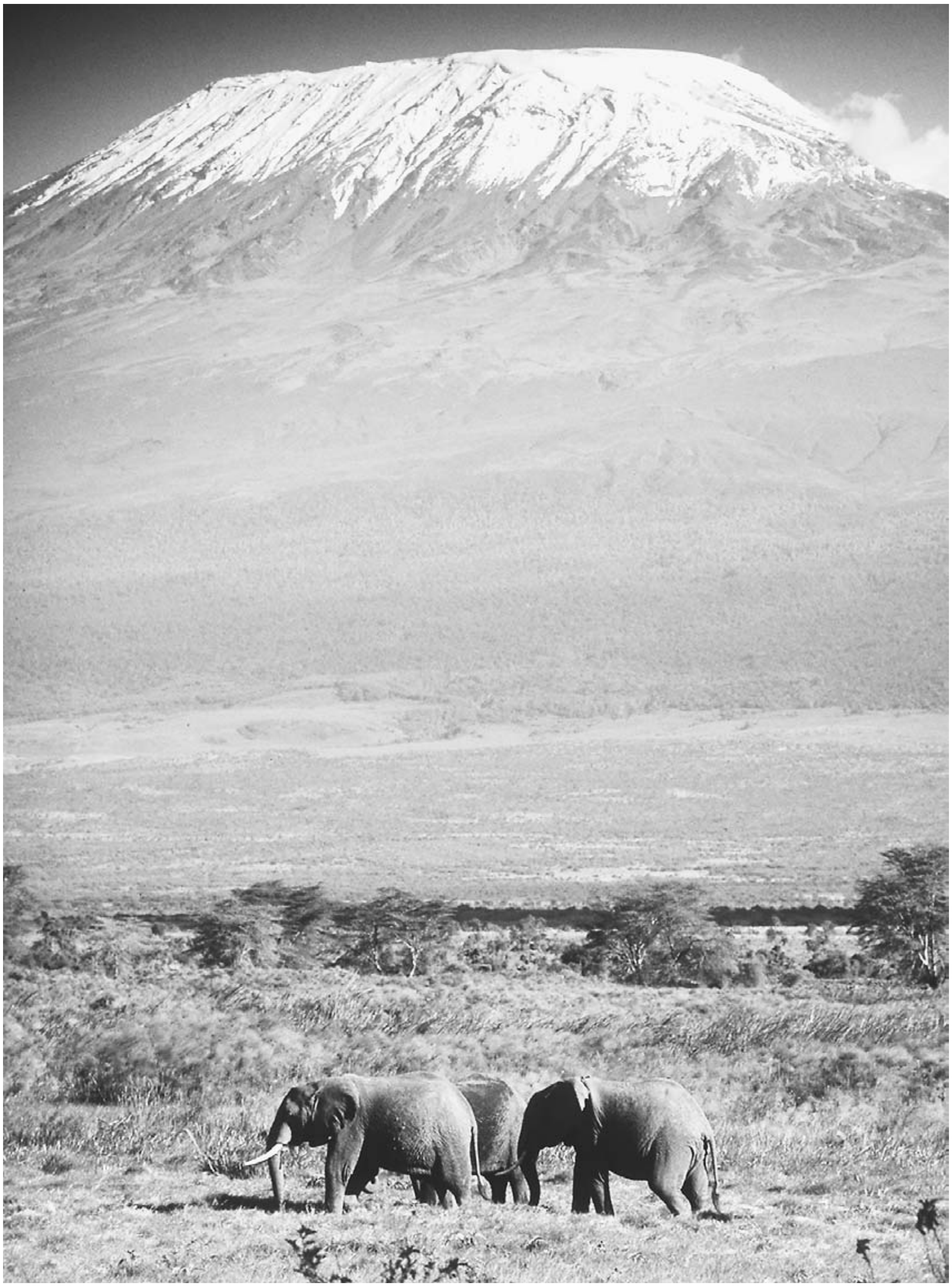
Earth is constantly moving, driven by forces beneath its surface. The interior of Earth itself is divided into three major sections: the crust, mantle, and core. The lithosphere is the upper layer of Earth's interior, including the crust and the brittle portion at the top of the mantle. Tectonism is the deformation of the lithosphere, and the term tectonics refers to the study of this deformation. Most notable among examples of tectonic deformation is mountain building, or orogenesis, discussed later in this essay.

The planet's crust is not all of one piece: it is composed of numerous plates, which are steadily moving in relation to one another. This movement is responsible for all manner of phenomena, including earthquakes, volcanoes, and mountain building. All these ideas and many more are encompassed in the concept of plate tectonics, which is the name for a branch of geologic and geophysical study and of a dominant principle often described as "the unifying theory of geology" (see Plate Tectonics).

CONTENTS UNDER PRESSURE. Tectonism results from the release and redistribution of energy from Earth's interior. This energy is either gravitational, and thus a function of the enormous mass at the planet's core, or thermal, resulting from the heat generated by radioactive decay. Differences in mass and heat within the planet's interior, known as pressure gradients, result in the deformation of rocks, placing many forms of stress and strain on them.

In scientific terms, *stress* is any attempt to deform an object, and strain is a change in dimension resulting from stress. Rocks experience stress in the form of tension, compression, and shear. Tension acts to stretch a material, whereas compression is a form of stress produced by the action of equal and opposite forces, whose effect is to reduce the length of a material. (Compression is a form of pressure.) Shear results from equal and opposite forces that do not act along the same plane. If a thick, hardbound book is lying flat, and one pushes the front cover from the side so that the covers and pages are no longer in alignment, is an example of shear.

Rocks manifest the strain resulting from these stresses by warping, sliding, or breaking. They may even flow, as though they were liquids, or melt and thus truly become liquid. As a result, Earth's interior may manifest faults, or fractures in rocks, as



MOUNTAINS SOMETIMES ARISE IN ISOLATION, AS WAS THE CASE WITH MOUNT KILIMANJARO IN TANZANIA. *(© T. Davis/Photo Researchers. Reproduced by permission.)*

well as folds, or bends in the rock structure. The effects can be seen on the surface in the form of subsidence, which is a depression in the crust; or uplift, the raising of crustal materials. Earthquakes and volcanic eruptions also may result.

OROGENESIS

There are two basic types of tectonism: epeirogenesis and orogenesis. The first takes its name from the Greek words *epeiros,* meaning "mainland," and *genesis,* or "origin." Epeirogenesis, which takes the form of either uplift or subsidence, is a chiefly vertical form of movement and plays little role in either plate tectonics or mountain building.

Orogenesis, on the other hand, *is* mountain building, as the prefix *oros* ("mountain") shows. Orogenesis involves the formation of mountain ranges by means of folding, faulting, and volcanic activity—lateral movements as opposed to vertical ones. Geologists typically use the term *orogenesis,* instead of just "mountain building," when discussing the formation of large belts of mountains from tectonic processes.

PLATE MARGINS. Plates may converge (move toward one another), diverge (move away from one another), or experience transform motion, meaning that they slide against one another. Convergence usually is associated with subduction, in which one plate is forced down into the mantle and eventually undergoes partial melting. This typically occurs in the ocean, creating a depression known as an oceanic trench.

There are three types of plate margins, or boundaries between plates, depending on the two types of crusts that interact: oceanic with oceanic, continental with continental, or continental with oceanic. Any of these margins may be involved in mountain formation. Orogenic belts, or mountain belts, typically are situated in subduction zones at convergent plate boundaries and consist of two types.

The first type occurs when igneous material (i.e., rock from volcanoes) forms on the upper plate of a subduction zone, causing the surface to rise. This can take place either in the oceanic crust, in which case the mountains formed are called island arcs, or along continental-oceanic margins. The Aleutian Islands are an example of an island arc, while the Andes range represent mountains formed by the subduction of an oceanic plate under a continental one.

The second type of mountain belt occurs when continental plates converge or collide. When continental plates converge, one plate may "try" to subduct the other, but ultimately the buoyancy of the lower plate (which floats, as it were, on the lithosphere) pushes it upward. The result is the creation of a wide, unusually thick or "tall" belt. An example is the Himalayas, the world's tallest mountain range, which is still being pushed upward as the result of a collision between India and Asia that happened some 30 million years ago. (See Plate Tectonics for more about continental drift and collisions between plates.)

REAL-LIFE APPLICATIONS

WHAT IS A MOUNTAIN?

In the 1995 film *The Englishman Who Went Up a Hill But Came Down a Mountain,* the British actor Hugh Grant plays an English cartographer, or mapmaker, sent in 1917 by his government to measure what is purportedly "the first mountain inside Wales." He quickly determines that according to standards approved by His Majesty, the "mountain" in question is, in fact, a hill. Much of the film's plot thereafter revolves around attempts on the part of the villagers to rescue their beloved mountain from denigration as a "hill," a fate they prevent by piling enough rocks and dirt onto the top to make it meet specifications.

This comedy aptly illustrates the somewhat arbitrary standards by which people define mountains. The British naturalist Roderick Peattie (1891–1955), in his 1936 book *Mountain Geography,* maintained that mountains are distinguished by their impressive appearance, their individuality, and their impact on the human imagination. This sort of qualitative definition, while it is certainly intriguing, is of little value to science; fortunately, however, more quantitative standards exist.

In Britain and the United States, a mountain typically is defined as a landform with an elevation of 985 ft. (300 m) above sea level. This was the standard applied in *The Englishman,* but the Welsh villagers would have had a hard time raising their "hill" to meet the standards used in continental Europe: 2,950 ft. (900 m) above sea level. This seems to be a more useful standard, because the British and American one would take in high plains and other nonmountainous regions of relatively great altitude. On the other hand, there are landforms in Scotland that rise only a few hundred meters above sea level, but their morphologic characteristics or shape seem to qualify them as mountains. Not only are their slopes steep, but the presence of glaciers and snow-capped peaks, with their attendant severe weather and rocky, inhospitable soil, also seem to indicate the topography associated with mountains.

Mountain Geomorphology

One area of the geologic sciences especially concerned with the study of mountains is geomorphology, devoted to the investigation of landforms. Geomorphologists studying mountains must draw on a wide variety of disciplines, including geology, climatology, biology, hydrology, and even anthropology, because, as discussed at the conclusion of this essay, mountains have played a significant role in the shaping of human social groups.

From the standpoint of geology and plate tectonics, mountain geomorphology embraces a complex of characteristic formations, not all of which are necessarily present in a given orogen, or mountain. These include forelands and foredeeps along the plains; foreland fold-and-thrust belts, which more or less correspond to "foothills" in layperson's terminology; and a crystalline core zone, composed of several types of rock, that is the mountain itself.

ENVIRONMENTAL ZONES. Mountain geomorphology classifies various environmental zones, from lowest to highest altitude. Near the bottom are flood plains, river terraces, and alluvial fans, all areas heavily affected by rivers flowing from higher elevations. (In fact, many of the world's greatest rivers flow from mountains, examples being the Himalayan Ganges and Indus rivers in Asia and the Andean Amazon in South America.) Farming villages may be found as high as the 9,845-ft. to 13,125-ft. range (3,000–4,000 m), an area known as a submontane, or forested region.

The tree line typically lies at an altitude of 14,765 ft. (4,500 m). Above this point, there is little human activity but plenty of geologic activity, including rock slides, glacial flow, and, at very high altitudes, avalanches. From the tree line upward, the altitude levels that mark a particular region are differentiated for the Arctic and tropical zones, with much lower altitudes in the Arctic mountains. For instance, the tree line lies at about 330 ft. (100 m) in the much colder Arctic zone.

Above the tree line is the subalpine, or montane, region. The mean slope angle of the mountain is less steep here than it is at lower or higher elevations: in the submontane, or forested region, below the tree line, the slope is about 30°, and above the subalpine, in the high alpine, the slope can become as sharp at 65°. In the subalpine, however, it is only about 20°, and because grass (if not trees) grows in this region, it is suited for grazing.

It may seem surprising to hear of shepherds bringing sheep to graze at altitudes of 16,400 ft. (5,000 m), as occurs in tropical zones. This does not necessarily mean that people live at such altitudes; more often than not, mountain dwellers have their settlements at lower elevations, and shepherds simply take their flocks up into the heights for grazing. Yet the ancient Bolivian city of Tiahuanaco, which flourished in about A.D. 600—some four centuries before the rise of the Inca—lay at an almost inconceivable altitude of 13,125 ft. (4,000 m), or about 2.5 times the elevation of Denver, Colorado, America's Mile-High City.

Classifying Mountains

There are several ways to classify mountains and groups of mountains. Mountain belts, as described earlier, typically are grouped according to formation process and types of plates: island arcs, continental arcs (formed with the subduction of an oceanic plate by a continental plate), and collisional mountain belts. Sometimes a mountain arises in isolation, an example being Kilimanjaro in Tanzania, Africa. Another example is Stone Mountain outside Atlanta, an exposed pluton, or a mass of crystalline igneous rock that forms deep in Earth's crust and rises. Many volcanoes, which we discuss later, arise individually, but mountains are most likely to appear in conjunction with other mountains. One such grouping, though far from the only one, is a mountain range, which can be defined as a relatively localized series of peaks and ridges.

RANGES, CHAINS, AND MASSES. Some of the world's most famous mountain ranges include the Himalayas, Karakoram Range, and Pamirs in central Asia; the Alps and Urals in Europe; the Atlas Mountains in Africa; the Andes in South America; and the Cascade Range, Sierra Nevada, Rocky Mountains, and Appalachians as well as their associated ranges in North America. Ranges affiliated with the Appalachians, for instance, include the Great Smokies in the south and the Adirondacks, Alleghenies, and Poconos in the north.

Several of the examples given here illustrate the fact that ranges are not the largest groupings of mountains. Sometimes series of ranges stretch across a continent for great distances in what are called mountain chains, an example of which is the Mediterranean chain of Balkans, Apennines, and Pyrenees that stretches across southern Europe.

There also may be irregular groupings of mountains, which lack the broad linear sweep of mountain ranges or chains and which are known as mountain masses. The mountains surrounding the Tibetan plateau represent an example of a mountain mass. Finally, ranges, chains, and masses of mountains may be combined to form vast mountain systems. An impressive example is the Alpine-Himalayan system, which unites parts of the Eurasian, Arabian, African, and Indo-Australian continental plates.

OTHER TYPES OF MOUNTAIN. There are certain special types of orogeny, as when ocean crust subducts continental crust—something that is not supposed to happen but occasionally does. This rare variety of subduction is called obduction, and the mountains produced are called ophiolites. Examples include the uplands near Troodos in Cyprus and the Taconic Mountains in upstate New York.

Fault-block mountains appear when two continental masses push against each other and the upper portion of a continental plate splits from the deeper rocks. A portion of the upper crust, usually several miles thick, begins to move slowly across the continent. Ultimately it runs into another mass, creating a ramp. This can result in unusually singular mountains, such as Chief Mountain in Montana, which slid across open prairie on a thrust sheet.

Under the ocean is the longest mountain chain on Earth, the mid-ocean ridge system, which runs down the center of the Atlantic Ocean and continues through the Indian and Pacific oceans. Lava continuously erupts along this ridge, releasing geothermal energy and opening up new strips of ocean floor. This brings us to a special kind of mountain, typically resulting from the sort of dramatic plate tectonic processes that also produce earthquakes: volcanoes.

Volcanoes

Most volcanoes are mountains, and for this reason, it is appropriate to discuss them together; however, a volcano is not necessarily a mountain. A volcano may be defined as a natural opening in Earth's surface through which molten (liquid), solid, and gaseous material erupts. The word *volcano* also is used to describe the cone of erupted material that builds up around the opening or fissure. Because these cones are often quite impressive in height, they frequently are associated with mountains.

The Popocatepetl volcano erupts, spewing ash, rocks, and gases. *(© Wesley Bocxe/Photo Researchers. Reproduced by permission.)*

Though volcanic activity has been the case of death and destruction, it is essential to the planet's survival. Volcanic activity is the principal process through which chemical elements, minerals, and other compounds from Earth's interior reach its surface. These substances, such as carbon dioxide, have played a major role in the development of the planet's atmosphere, waters, and soils. Even today, soil in volcanic areas is among the richest on Earth. Volcanoes provide additional benefits in their release of geothermal energy, used for heating and other purposes in such countries as Iceland, Italy, Hungary, and New Zealand (see Energy and Earth). In addition, volcanic activity beneath the oceans promises to supply almost limitless geothermal energy, once the technology for its extraction becomes available.

FORMATION OF VOLCANOES. As noted earlier, land volcanoes are formed in coastal areas where continental and oceanic plates converge. As the oceanic plate is subducted and pushed farther and farther beneath the continental surface, the buildup of heat and pressure results in the melting of rock. This molten rock, or magma, tends to rise toward the surface and collect in magma reservoirs. Pressure buildup in the magma reservoir ultimately pushes the magma upward through cracks in Earth's crust, creating a volcano.

Volcanoes also form underwater, in which case they are called seamounts. Convergence of oceanic plates causes one plate to sink beneath the other, creating an oceanic trench; as a result, magma rises from the subducted plate to fashion volcanoes. If the plates diverge, magma seeps upward at the ridge or margin between plates, producing more seafloor. This process, known as seafloor spreading, leads to the creation of volcanoes on either side of the ridge.

In some places a plate slides over a stationary area of volcanic activity, known as a hot spot. These are extremely hot plumes of magma that well up from the crust, though not on the edge or margin of a plate. A tectonic plate simply drifts across the hot spot, and as it does, the area just above the hot spot experiences volcanic activity. Hot spots exist in Hawaii, Iceland, Samoa, Bermuda, and America's Yellowstone National Park.

CLASSIFYING VOLCANOES. Volcanoes can be classified in terms of their volcanic activity, in which case they are labeled as active (currently erupting), dormant (not currently erupting but likely to do so in the future), or extinct. In the case of an extinct volcano, no eruption has been noted in recorded history, and it is likely that the volcano has ceased to erupt permanently.

In terms of shape, volcanoes fall into four categories: cinder cones, composite cones, shield volcanoes, and lava domes. These types are distinguished not only by morphologic characteristics but also by typical sizes and even angles of slope. For instance, cinder cones, built of lava fragments, have slopes of 30° to 40°, and are seldom more than 1,640 ft. (500 m) in height.

Composite cones, or stratovolcanoes, are made up of alternating layers of lava (cooled magma), ash, and rock. (The prefix *strato* refers to these layers.) They may slope as little as 5° at the base and as much as 30° at the summit. Stratovolcanoes may grow to be as tall as 2–3 mi. (3.2–4.8 km) before collapsing and are characterized by a sharp, dramatic shape. Examples include Fuji, a revered mountain that often serves as a symbol of Japan, and Washington state's Mount Saint Helens.

A shield volcano, which may be a solitary formation and often is located over a hot spot, is built from lava flows that pile one on top of another. With a slope as little as 2° at the base and no more than 10° at the summit, shield volcanoes are much wider than stratovolcanoes, but sometimes they can be impressively tall. Such is the case with Mauna Loa in Hawaii, which at 13,680 ft. (4,170 m) above sea level is the world's largest active volcano. Likewise, Mount Kilimanjaro, though long ago gone dormant, is the tallest mountain in Africa.

Finally, there are lava domes, which are made of solid lava that has been pushed upward. Closely related is a volcanic neck, which often forms from a cinder cone. In the case of a volcanic neck, lava rises and erupts, leaving a mountain that looks like a giant gravel heap. Once it has become extinct, the lava inside the volcano begins to solidify. Over time the rock on the exterior wears away, leaving only a vent filled with solidified lava, usually in a funnel shape. A dramatic example of this appears at Shiprock, New Mexico.

VOLCANIC ERUPTIONS. Volcanoes frequently are classified by the different ways in which they erupt. These types of eruption, in turn, result from differences in the material being disgorged from the volcano. When the magma is low in gas and silica (silicon dioxide, found in sand and rocks), the volcano erupts in a relatively gentle way. Its lava is thin and spreads quickly. Gas and silica–rich magma, on the other hand, brings about a violent explosion that yields tarlike magma.

There are four basic forms or phases of volcanic eruption: Hawaiian, Strombolian, Vulcanian, and Peleean. The Hawaiian phase is simply a fountain-like gush of runny lava, without any explosions. The Strombolian phase (named after a volcano on a small island off the Italian penin-

Crater Lake in Oregon is the result of the collapse of a magma chamber after a volcanic eruption. This collapse forms a bowl-like crater called a caldera, which fills with water. *(© Francois Gohier/Photo Researchers. Reproduced by permission.)*

sula) involves thick lava and mild explosions. In a Vulcanian phase, magma has blocked the volcanic vent, and only after an explosion is the magma released, with the result that tons of solid material and gases are hurled into the sky. Most violent of all is the Peleean, named after Mount Pelée on Martinique in the Caribbean (discussed later). In the Peleean phase, the volcano disgorges thick lava, clouds of gas, and fine ash, all at formidable velocities.

Accompanying a volcanic eruption in many cases are fierce rains, the result of the expulsion of steam from the volcano, after which the steam condenses in the atmosphere to form clouds. Gases thrown into the atmosphere are often volatile and may include hydrogen sulfide, fluorine, carbon dioxide, and radon. All are detrimental to human beings when present in sufficient quantity, and radon is radioactive.

Not surprisingly, the eruption of a volcano completely changes the morphologic characteristics of the landform. During the eruption a crater is formed, and out of this flows magma and ash, which cool to form the cone. In some cases, the magma chamber collapses just after the eruption, forming a caldera, or a large, bowl-shaped crater. These caldera (the plural as well as singular form) may fill with water, as was the case at Oregon's Crater Lake.

INFAMOUS VOLCANIC DISASTERS. Volcanoes result from some of the same tectonic forces as earthquakes (see Seismology), and, not surprisingly, they often have resulted in enormous death and destruction. Some remarkable examples include:

- *Vesuvius, Italy, A.D. 79 and 1631:* Situated along the Bay of Naples in southern Italy, Vesuvius has erupted more than 50 times during the past two millennia. Its most famous eruption occurred in A.D. 79, when the Roman Empire was near the height of its power. The first-century eruption buried the nearby towns of Pompeii and Herculaneum, where bodies and buildings were preserved virtually intact until excavation of the area in 1748. Another eruption, in 1631, killed some 4,000 people.
- *Krakatau, Indonesia, A.D. 535 (?) and 1883:* The most famous eruption of Krakatau occurred in 1883, resulting in the loss of some 36,000 lives. The explosion, which was heard 3,000 mi. away, threw 70-lb. (32-kg) boulders as far as 50 mi. (80 km). It also produced a tsunami, or tidal wave, 130 ft.

(40 m) high, which swept away whole villages. In addition, the blast hurled so much dust into the atmosphere that the Moon appeared blue or green for two years. It is also possible that Krakatau erupted in about A.D. 535, causing such a change in the atmosphere that wide areas of the world experienced years without summer. (See Earth Systems for more on this subject.)

- *Tambora, Indonesia, 1815:* Another Indonesian volcano, Tambora, killed 12,000 people when it erupted in 1815. As with Krakatau in 535, this eruption was responsible for a year without summer in 1816 (see Earth Systems).
- *Pelée, Martinique, 1902:* When Mount Pelée erupted on the Caribbean island of Martinique, it sent tons of poisonous gas and hot ash spilling over the town of Saint-Pierre, killing all but four of its 29,937 residents.
- *Saint Helens, Washington, 1980:* Relatively small compared with earlier volcanoes, the Mount Saint Helens blast is still significant because it was so recent and took place in the United States. The eruption sent debris flying upward 1,300 ft. (396 m) and caused darkness over towns as far as 85 mi. (137 km) away. Fifty-seven people died in the eruption and its aftermath.
- *Pinatubo, Philippines, 1991:* Dormant for 600 years, Mount Pinatubo began to rumble one day in 1991 and, after a few days, erupted in a cloud that spread ash 6 ft. (1.83 m) deep along a radius of 2 mi. (3.2 km). A U.S. air base 15 mi. (24 km) away was buried. The blast threw 20 million tons (18,144,000 metric tons) of sulfuric acid 12 mi. (19 km) into the stratosphere, and the cloud ultimately covered the entire planet, resulting in moderate cooling for a few weeks.

THE IMPACT OF MOUNTAINS

Volcanic eruptions are among the most dramatic effects produced by mountains, but they are far from the only ones. Every bit as fascinating are the effects mountains produce on the weather, on the evolution of species, and on human society. In each case, mountains serve as a barrier or separator—between masses of air, clouds, and populations.

Wind pushes air and moisture-filled clouds up mountain slopes, and as the altitude increases, the pressure decreases. As a result, masses of warm, moist air become larger, cooler, and less dense. This phenomenon is known as *adiabatic expansion,* and it is the same thing that happens when an aerosol can is shaken, reducing the pressure of gases inside and cooling the surface of the can. Under the relatively high-pressure and high-temperature conditions of the flatlands, water exists as a gas, but in the heights of the mountaintops, it cools and condenses, forming clouds.

RAIN SHADOWS. As the clouds rise along the side of the mountain, they begin to release heavy droplets in the form of rain and, at higher altitudes, snow. By the time the cloud crosses the top of the mountain, however, it will have released most of its moisture, and hence the other side of the mountain may be arid. The leeward side, or the side opposite the wind, becomes what is called a rain shadow.

Although they are only 282 mi. (454 km) apart, the cities of Seattle and Spokane, Washington, have radically different weather patterns. Famous for its almost constant rain, Seattle lies on the windward, or wind-facing, side of the Cascade Range, toward the Pacific Ocean. On the leeward side of the Cascades is Spokane, where the weather is typically warm and dry. Though it is only on the other side of the state, Spokane might as well be on the other side of the continent. Indeed, it is associated more closely with the arid expanses of Idaho, whereas Seattle belongs to a stretch of cold, wet Pacific terrain that includes San Francisco and Portland, Oregon.

Much of the western United States consists of deserts formed by rain shadows or, in some cases, double rain shadows. Much of New Mexico, for instance, lies in a double rain shadow created by the Rockies in the west and Mexico's Sierra Madres to the south. In southern California, tall redwoods line the lush windward side of the Sierra Nevadas, while Death Valley and the rest of the Mojave Desert lies in the rain shadow on the eastern side. The Great Basin that covers eastern Oregon, southern Idaho, much of Utah, and almost all of Nevada, likewise is created by the rain shadow of the Sierra Nevada-Cascade chain.

MOUNTAINS AND SPECIES. One of the most intriguing subjects involved in the study of mountains is their effects on large

KEY TERMS

ACTIVE: A term to describe a volcano that is currently erupting.

COMPRESSION: A form of stress produced by the action of equal and opposite forces, the effect of which is to reduce the length of a material. Compression is a form of pressure.

CONVERGENCE: A tectonic process whereby plates move toward each other.

CRUST: The uppermost division of the solid earth, representing less than 1% of its volume and varying in depth from 3 mi. to 37 mi. (5–60 km). Below the crust is the mantle.

DIVERGENCE: A tectonic process whereby plates move away from each other.

DORMANT: A term to describe a volcano that is not currently erupting but is likely to do so in the future.

EPEIROGENESIS: One of two principal forms of tectonism, the other being orogenesis. Derived from the Greek words *epeiros* ("mainland") and *genesis* ("origins"), epeirogenesis takes the form of either uplift or subsidence.

EXTINCT: A term to describe a volcano for which no eruption has been known in recorded history. In this case, it is likely that the volcano has ceased to erupt permanently.

GEOMORPHOLOGY: An area of physical geology concerned with the study of landforms, with the forces and processes that have shaped them, and with the description and classification of various physical features on Earth.

HOT SPOT: A region of high volcanic activity.

LANDFORM: A notable topographical feature, such as a mountain, plateau, or valley.

LITHOSPHERE: The upper layer of Earth's interior, including the crust and the brittle portion at the top of the mantle.

MANTLE: The thick, dense layer of rock, approximately 1,429 mi. (2,300 km) thick, between Earth's crust and its core.

MORPHOLOGY: Structure or form or the study thereof.

MOUNTAIN CHAIN: A series of ranges stretching across a continent for a great distance.

MOUNTAIN MASS: An irregular grouping of mountains, which lacks the broad linear sweep of a range or chain.

MOUNTAIN RANGE: A relatively localized series of peaks and ridges.

MOUNTAIN SYSTEM: A combination of ranges, chains, and masses of mountains that stretches across vast distances, usually encompassing more than one continent.

OROS: A Greek word meaning "mountain," which appears in such words as *orogeny,* a variant of *orogenesis; orogen,* another term for "mountain" and *orogenic,* as in "orogenic belt."

OROGENESIS: One of two principal forms of tectonism, the other being epeirogenesis. Derived from the Greek words *oros* ("mountain") and *genesis* ("origin"), orogenesis involves the formation of mountain ranges by means of folding, faulting, and volcanic activity. The processes of orogenesis play a major role in plate tectonics.

PLATE MARGINS: Boundaries between plates.

KEY TERMS CONTINUED

PLATE TECTONICS: The name both of a theory and of a specialization of tectonics. As an area of study, plate tectonics deals with the large features of the lithosphere and the forces that shape them. As a theory, it explains the processes that have shaped Earth in terms of plates and their movement.

PLATES: Large, movable segments of the lithosphere.

SHEAR: A form of stress resulting from equal and opposite forces that do not act along the same line. If a thick, hardbound book is lying flat, and one pushes the front cover from the side so that the covers and pages are no longer aligned, this is an example of shear.

STRAIN: The ratio between the change in dimension experienced by an object that has been subjected to stress and the original dimensions of the object.

STRESS: In general terms, any attempt to deform a solid. Types of stress include tension, compression, and shear.

SUBSIDENCE: A term that refers either to the process of subsiding, on the part of air or solid earth, or, in the case of solid earth, to the resulting formation. Subsidence thus is defined variously as the downward movement of air, the sinking of ground, or a depression in Earth's crust.

TECTONICS: The study of tectonism, including its causes and effects, most notably mountain building.

TECTONISM: The deformation of the lithosphere.

TENSION: A form of stress produced by a force that acts to stretch a material.

TOPOGRAPHY: The configuration of Earth's surface, including its relief as well as the position of physical features.

UPLIFT: A process whereby the surface of Earth rises, as the result of either a decrease in downward force or an increase in upward force.

VOLCANO: A natural opening in Earth's surface through which molten (liquid), solid, and gaseous material erupts. The word volcano is also used to describe the cone of erupted material that builds up around the opening or fissure.

groups of plants, animals, and humans. Mountains may separate entire species, creating pockets of flora and fauna virtually unknown to the rest of the world. Thus, during the 1990s, huge numbers of species that had never been catalogued were discovered in the mountains of southeast Asia.

The formation of mountains and other landforms may even lead to speciation, a phenomenon in which members of a species become incapable of reproducing with other members, thus creating a new species. When the Colorado River cut open the Grand Canyon, it separated groups of squirrels that lived in the high-altitude pine forest. Over time these populations ceased to interbreed, and today the Kaibab squirrel of the north rim and the Abert squirrel of the south are separate species, no more capable of interbreeding than humans and apes.

HUMAN SOCIETIES AND MOUNTAINS. Although the Appalachians of the eastern United States are hundreds of millions of years old, most ranges are much younger. Most will erode or otherwise cease to exist in a relatively short time (short, that is, by geologic standards), yet to humans throughout the ages, mountains have seemed a symbol of permanence. This is just one aspect of mountains' impact on the human psyche.

In his 1975 study of symbolism in political movements, *Utopia and Revolution,* Melvin J. Lasky devoted considerable space to the mountain and its association with divinity through figures such as the Greek Olympians and Noah and Moses in the Bible. Clearly, mountains have proved enormously influential on human attitudes, and nowhere is this more obvious than in relation to the people who live *in* the mountains. Whether the person is a coal miner from Appalachia or a rancher from the Rockies, a Scottish highlander or a Quechua-speaking Peruvian, the mentality is similar, characterized by a combination of hardiness, fierce independence, and disdain for lowland ways.

These characteristics, combined with the harsh weather of the mountains, have made mountain warfare a challenge to lowland invaders. This explains the fact that Switzerland has kept itself free from involvement in European wars since Napoleon's time, and why the independent Scottish Highlands were long a thorn in England's side. It also explains why neither the British nor the Russian empires could manage to control Afghanistan fully during their struggle over that mountainous nation in the late nineteenth and early twentieth centuries.

Britain eventually pulled out of the "Great Game," as this struggle was called, but Russia never really did. Many years later, the Soviets became bogged down in a war in Afghanistan that they could not win. The war, which lasted from 1979 to 1989, helped bring about the end of the Soviet Union and its system of satellite dictatorships. More than a decade later, as the United States launched strikes against Afghanistan in 2001, a superpower once again faced the challenge posed by one of the poorest, most inhospitable nations on Earth.

But the independence of the mountaineer is deceptive; in fact, mountains have little to offer, economically, other than their beauty and the resources deep beneath their surfaces. In other words, they are really of value only to flatland tourists and mining companies. Since few mountain environments offer much promise agriculturally, the people of the mountains are dependent on the flatlands for sustenance. Gorgeous and rugged as they are, such mountainous states as Colorado or Wyoming might be as poor as Afghanistan were it not for the fact that they belong to a larger political unit, the United States.

WHERE TO LEARN MORE

A Geological History of Rib Mountain, Wisconsin (Web site). <http://www.uwmc.uwc.edu/geography/ribmtn/ribmtn.htm>.

Gore, Pamela. "Physical Geology at Georgia Perimeter College" (Web site). <http://www.gpc.peachnet.edu/~pgore/geology/geo101.htm>.

Kraulis, J. A., and John Gault. *The Rocky Mountains: Crest of a Continent.* New York: Facts on File, 1987.

Michigan Technological University Volcanoes Page (Web site). <http://www.geo.mtu.edu/volcanoes/>.

Oregon Geology—Cascade Mountains (Web site). <http://sarvis.dogami.state.or.us/learnmore/Cascades.HTM>.

Prager, Ellen J., Kate Hutton, Costas Synolakis, et al. *Furious Earth: The Science and Nature of Earthquakes, Volcanoes, and Tsunamis.* New York: McGraw-Hill, 2000.

Schaer, Jean-Paul, and John Rodgers. *The Anatomy of Mountain Ranges.* Princeton, NJ: Princeton University Press, 1987.

Sigurdsson, Haraldur. *Encyclopedia of Volcanoes.* San Diego: Academic Press, 2000.

Silver, Donald M., and Patricia Wynne. *Earth: The Ever-Changing Planet.* New York: Random House, 1989.

Volcanoes, Glaciers, and Plate Tectonics: The Geology of the Mono Basin (Web site). <http://www.r5.fs.fed.us/inyo/vvc/mono/vg&26pt.htm>.

EROSION

CONCEPT

Erosion is a broadly defined group of processes involving the movement of soil and rock. This movement is often the result of flowing agents, whether wind, water, or ice, which sometimes behaves like a fluid in the large mass of a glacier. Gravitational pull may also influence erosion. Thus, erosion, as a concept in the earth sciences, overlaps with mass wasting or mass movement, the transfer of earth material down slopes as a result of gravitational force. Even more closely related to erosion is weathering, the breakdown of rocks and minerals at or near the surface of Earth owing to physical, chemical, or biological processes. Some definitions of erosion even include weathering as an erosive process. Though most widely known as a by-product of irresponsible land use by humans and for its negative effect on landforms, erosion is neither unnatural nor without benefit. Far more erosion occurs naturally than as a result of land development, and a combination of weathering and erosion is responsible for producing the soil from which Earth's plants grow.

HOW IT WORKS

Weathering

The first step in the process of erosion is weathering. Weathering, in a general sense, occurs everywhere: paint peels; metal oxidizes, resulting in its tarnishing or rusting; and any number of products, from shoes to houses, begin to show the effects of physical wear and tear. The scuffing of a shoe, cracks in a sidewalk, or the chipping of glass in a gravel-spattered windshield are all examples of physical weathering. On the other hand, the peeling of paint is usually the result of chemical changes, which have reduced the adhesive quality of the paint. Certainly oxidation is a chemical change, meaning that it has not simply altered the external properties of the item but also has brought about a change in the way that the atoms are bonded.

Weathering, as the term is used in the geologic sciences, refers to these and other types of physical and chemical changes in rocks and minerals at or near the surface of Earth. A mineral is a substance that occurs naturally and is usually inorganic, meaning that it contains carbon in a form other than that of an oxide or a carbonate, neither of which is considered organic. It typically has a crystalline structure, or one in which the constituent parts have a simple and definite geometric arrangement repeated in all directions. Rocks are simply aggregates or combinations of minerals or organic material or both.

TWO AND ONE-HALF KINDS OF WEATHERING. There are three kinds of weathering (or perhaps two and one-half, since the third incorporates aspects of the first two): physical or mechanical, chemical, and biological. Physical or mechanical weathering takes place as a result of such factors as gravity, friction, temperature, and moisture. Gravity may cause a rock to drop from a height, such that it falls to the ground and breaks into pieces, while the friction of wind-borne sand may wear down a rock surface. Changes in temperature and moisture cause expansion and contraction of materials, as when water seeps into a crack in a rock and then freezes, expanding and splitting the rock.

Nicola River Canyon in British Columbia shows the effects of freeze-thaw and erosion by wind and rain. *(© K. Svensson/Photo Researchers. Reproduced by permission.)*

Minerals are chemical compounds; thus, whereas physical weathering attacks the rock as a whole, chemical weathering effects the breakdown of the minerals that make up the rock. This breakdown may lead to the dissolution of the minerals, which then are washed away by water or wind or both, or it may be merely a matter of breaking the minerals down into simpler compounds. Reactions that play a part in this breakdown may include oxidation, mentioned earlier, as well as carbonation, hydrolysis (a reaction with water that results in the separation of a compound to form a new substance or substances), and acid reactions. For instance, if coal has been burned in an area, sulfur impurities in the air react with water vapor (an example of hydrolysis) to produce acid rain, which can eat away at rocks. Rainwater itself is a weak acid, and over the years it slowly dissolves the marble of headstones in old cemeteries.

As noted earlier, there are either three or two and one-half kinds of weathering, depending on whether one considers biological weathering a third variety or merely a subset of physical and chemical weathering. The weathering exerted by organisms (usually plants rather than animals) on rocks and minerals is indeed chemical and physical, but because of the special circumstances, it is useful to consider it individually. There is likely to be a long-term interaction between the organism and the geologic item, an obvious example being a piece of moss that grows on a rock. Over time, the moss will influence both physical and chemical weathering through its attendant moisture as well as its specific chemical properties, which induce decomposition of the rock's minerals.

Unconsolidated Material

The product of weathering in rocks or minerals is unconsolidated, meaning that it is in pieces, like gravel, though much less uniform in size. This is called regolith, a general term that describes a layer of weathered material that rests atop bedrock. Sand and soil, including soil mixed with loose rocks, are examples of regolith. Regolith is, in turn, a type of sediment, material deposited at or near Earth's surface from a number of sources, most notably preexisting rock.

Every variety of unconsolidated material has its own angle of repose, or the maximum angle at which it can remain standing. Piles of rocks may have an angle of repose as high as 45°, whereas dry sand has an angle of only 34°. The addition of water can increase the angle of repose, as anyone who has ever strengthened a sand castle by adding water to it knows. Suppose one builds a sand castle in the morning, sloping the sand at angles that would be impossible if it were dry. By afternoon, as wind and sunlight dry out the sand, the sand castle begins to fall apart, because its angle of repose is too high for the dry sand.

Water gives sand surface tension, the same property that causes water that has been spilled on a table to bead up rather than lie flat. If too much water is added to the sand, however, the sand becomes saturated and will flow, a process called lateral spreading. On the other hand, with too little moisture, the material is susceptible to erosion. Unconsolidated material in nature generally has a slope less than its angle of repose, owing to the influence of wind and other erosive forces.

INTRODUCTION TO MASS WASTING

There are three general processes whereby a piece of earth material can be moved from a high outcropping to the sea: weathering, mass wasting, and erosion. In the present context, we are concerned primarily with the last of these processes, of course, and secondarily with weathering, inasmuch as it contributes to erosion. A few words should be said about mass wasting, however, which, in its slower forms (most notably, creep), is related closely to erosion.

Mechanical or chemical processes, or a combination of the two, acting on a rock to dislodge it from a larger sample (e.g., separating a rock from a boulder) is an example of weathering, as we have seen. If the pieces of rock are swept away by a river in a valley below the outcropping, or if small pieces of rock are worn away by high winds, the process is erosion. Between the outcropping and the river below, if a rock has been broken apart by weathering, it may be moved farther along by mass-wasting processes, such as creep or fall.

REAL-LIFE APPLICATIONS

MASS WASTING IN ACTION

One of the principal sources of erosion is gravity, which is also the force behind creep, the slow downward movement of regolith along a hill slope. The regolith begins in a condition of unstable equilibrium, like a soda can lying on its side rather than perpendicular to a table's surface: in both cases, the object remains in place, yet a relatively small disturbance would be enough to dislodge it.

Changes in temperature or moisture are among the leading factors that result in creep. A variation in either can cause material to expand or contract, and freezing or thawing may be enough to shake regolith from its position of unstable equilibrium. Water also can provide lubrication, or additional weight, that assists the material in moving. Though it is slow, over time creep can produce some of the most dramatic results of any mass-wasting process. It can curve tree trunks at the base, break or dislodge retaining walls, and overturn objects ranging from fence posts to utility poles to tombstones.

OTHER VARIETIES OF FLOW. Creep is related to another slow mass-wasting process, known as solifluction, that occurs in the active layer of permafrost—that is, the layer that thaws in the summertime. The principal difference between creep and solifluction is not the speed at which they take place (neither moves any faster than about 0.5 in. [1 cm] per year) but the materials involved. Both are examples of flow, a chaotic form of mass wasting in which masses of material that are not uniform move downslope. With the exception of creep and solifluction, most forms of flow are comparatively rapid, and some are extremely so.

Because it involves mostly dry material, creep is an example of granular flow, which is composed of 0% to 20% water; on the other hand, solifluction, because of the ice component, is an instance of slurry flow, consisting of 20% to 40% water. If the water content is more than 40%, a slurry flow is considered a stream. Types of granular flow that move faster than creep range from earth flow to debris avalanche. Both earth flow and debris flow, its equivalent in slurry form, move at a broad range of speeds, anywhere from about 4 in. (10 cm) per year to 0.6 mi. (1 km) per *hour.* Grain flow can be as fast as 60 mi. (100 km) per hour, and mud flow is even faster. Fastest of all is debris avalanche, which may achieve speeds of 250 mi. (400 km) per hour.

OTHER TYPES OF MASS WASTING. Other varieties of mass wasting include slump, slide, and fall. Slump occurs when a mass of regolith slides over or creates a concave surface

Rock arches formed by erosion. *(© N. R. Rowan/Photo Researchers. Reproduced by permission.)*

(one shaped like the inside of a bowl.) The result is the formation of a small, crescent-shaped cliff, known as a scarp, at the upper end—rather like the crest of a wave. Slump often is classified as a variety of slide, in which material moves downhill in a fairly coherent mass (i.e., more or less in a section or group) along a flat or planar surface. These movements are sometimes called rock slides, debris slides, or, in common parlance, landslides.

In contrast to most other forms of mass wasting, in which there is movement along slopes that are considerably less than 90°, fall occurs at angles almost perpendicular to the ground. The "Watch for Falling Rock" signs on mountain roads may be frightening, and rock or debris fall is certainly one of the more dramatic forms of mass wasting. Yet the variety of mass wasting that has the most widespread effects on the morphology or shape of landforms is the slowest one—creep. (For more about the varieties of mass wasting, see Mass Wasting.)

WHAT CAUSES EROSION?

As noted earlier, the influences behind erosion are typically either gravity or flowing media: water, wind, and even ice in glaciers. Liquid water is the substance perhaps most readily associated with erosion. Given enough time, water can wear away just about anything, as proved by the carving of the Grand Canyon by the Colorado River.

Dubbed the *universal solvent* for its ability to dissolve other materials, water almost never appears in its pure form, because it is so likely to contain other substances. Even "pure" mountain water contains minerals and pieces of the rocks over which it has flowed, a testament to the power of water in etching out landforms bit by bit. Nor does it take a rushing mountain stream or crashing waves to bring about erosion; even a steady drip of water is enough to wear away granite over time.

MOVING WATER. Along coasts, pounding waves continually alter the shoreline. The sheer force of those walls of water, a result of the Moon's gravitational pull (and, to a lesser extent, the Sun's), is enough to wear away cliffs, let alone beaches. In addition, waves carry pieces of pebble, stone, and sand that cause weathering in rocks. Waves even can bring about small explosions in pockmarked rock surfaces by trapping air in small cracks; eventually the pressure becomes great enough that the air escapes, loosening pieces of the rock.

In addition to the erosive power of saltwater waves on the shore, there is the force exerted by running water in creeks, streams, and rivers. As the river moves, pushing along sediment and other materials eroded from the streambed or riverbed, it carves out deep chasms in the bedrock beneath. These moving bodies of water continually reshape the land, carrying soil and debris downslope, or from the source of the river to its mouth or delta. A delta is a region of sediment formed when a river enters a larger body of water, at which point the reduction in velocity on the part of the river current leads to the widespread deposition (depositing) of sediment. It is so named because its triangular shape resembles that of the Greek letter delta, Δ.

Water at the bottom of a large body, such as a pond or lake, also exerts erosive power. Then there is the influence of falling rain. Assuming ground is not protected by vegetation, raindrops can loosen particles of soil, sending them scattering in all directions. A rain that is heavy enough may dislodge whole layers of topsoil and send them rushing away in a swiftly moving current. The land left behind may be rutted and scarred, much of its best soil lost for good.

Just as erosion gives to the soil, it also can take away. Whereas erosion on the Nile delta acted to move rich, black soil into the region (hence, the ancient Egyptians' nickname for their country, the "black land"), erosion also can remove soil layers. As is often the case, it is much easier to destroy than to create: 1 in. (2.5 cm) of soil may take as long as 500 years to form, yet a single powerful rainstorm or windstorm can sweep it away.

GLACIERS

Ice, of course, is simply another form of water, but since it is solid, its physical (*not* its chemical) properties are quite different. Generally, physical sciences, such as physics or chemistry, treat as fluid all forms of matter that flow, whether they are liquid or gas. Normally, no solids are grouped under the heading of "fluid," but in the earth sciences there is at least one type of solid object that behaves as though it were fluid: a glacier.

A glacier is a large, typically moving mass of ice either on or adjacent to a land surface. It does not flow in the same way that water does; rather, it is moved by gravity, as a consequence of its extraordinary weight. Under certain conditions, a glacier may have a layer of melted water surrounding it, which greatly enhances it mobility. Regardless of whether it has this lubricant, however, a glacier steadily moves forward, carrying pieces of rock, soil, and vegetation with it.

These great rivers of ice gouge out pieces of bedrock from mountain slopes, fashioning deep valleys. Ice along the bottom of the glacier pulls away rocks and soil, which assist it in wearing

A "DUST DEVIL," OR A SMALL WHIRLWIND, CARRIES WITH IT DEBRIS AND SAND. *(© Clem Haagner/Photo Researchers. Reproduced by permission.)*

away bedrock. The fjords of Norway, where high cliffs surround narrow inlets whose depths extend many thousands of feet below sea level, are a testament to the power of glaciers in shaping the Earth. The fact that the fjords came into existence only in the past two million years, a product of glacial activity associated with the last ice age, is evidence of something else remarkable about glaciers: their speed.

"Speed," of course, is a relative term when speaking about processes involved in the shaping of the planet. A "fast" glacier, one whose movement is assisted by a wet and warm (again, *relatively* warm!) maritime climate, moves at the rate of about 980 ft. (300 m) per year. Examples include not only the glaciers that shaped the fjords, but also the active Franz Josef glacier in southern New Zealand. By contrast, in the dry, exceptionally cold, inland climate of Antarctica, the Meserve glacier moves at the rate of just 9.8 ft. (3 m) per year.

WIND

The erosion produced by wind often is referred to as an eolian process, the name being a reference to Aeolus, the Greek god of the winds encountered in Homer's *Odyssey* and elsewhere. Eolian processes include the erosion, transport, and deposition of earth material owing to the action of wind. It is most pronounced in areas that lack effective ground cover in the form of solidly rooted, prevalent vegetation.

Eolian erosion in some ways is less forceful than the erosive influence of water. Water, after all, can lift heavier and larger particles than can the winds. Wind, however, has a much greater frictional component in certain situations. This is particularly true when the wind carries sand, every grain of which is like a cutting tool. In some desert regions the bases of rocks or cliffs have been sandblasted, leaving a mushroom-shaped formation. The wind could not lift the fine grains of sand very high, but in places where it has been able to do its work, it has left an indelible mark.

The Dust Bowl and Human Contribution to Erosion

Though human actions are not a direct cause of erosion, human negligence or mismanagement often has prepared the way for erosive action by wind, water, or other agents. Interesting, soil itself, formed primarily by chemical weathering and enhanced by biological activity in the sediment, is a product of nature's erosive powers. Erosion transports materials from one place to another, robbing the soil in one place and greatly enhancing it in another.

This is particularly the case where river deltas are concerned. By transporting sediment and depositing it in the delta, the river creates an area of extremely fertile soil that, in some cases, has become literally the basis for civilizations. The earliest civilizations of the Western world, in Egypt and Sumer, arose in the deltas of the Nile and the Tigris-Euphrates river systems, respectively.

Erosion on the Great Plains. An extreme example of the negative effects on the soil that can come from erosion (and, ultimately, from human mismanagement) took place in Texas, Oklahoma, Colorado, and Kansas during the 1930s. In the preceding years, farmers unwittingly had prepared the way for vast erosion by overcultivating the land and not taking proper steps to preserve its moisture against drought. In some places farmers alternated between wheat cultivation and livestock grazing on particular plots of land.

The soil, already weakened by raising wheat, was damaged further by the hooves of livestock, and thus when a period of high winds began at the height of the Great Depression (1929–41), the land was particularly vulnerable. The winds carried dust to places as far away as the eastern seaboard, in some cases removing topsoil to a depth of 3–4 in. (7–10 cm). Dunes of dust as tall as 15–20 ft. (4.6–6.1 m) formed, and the economic blight of the Depression was compounded for the farmers of the plains states, many of whom lost everything.

Out of the Dust Bowl era came some of the greatest American works of art: the 1939 film *Wizard of Oz,* John Steinbeck's book *The Grapes of Wrath* and the acclaimed motion picture (1939 and 1940, respectively), as well as Dorothea Lange's haunting photographs of Dust Bowl victims. The Dust Bowl years also taught farmers and agricultural officials a lesson about land use, and in later years farming practices changed. Instead of alternating one year of wheat growing with one year in which a field lay fallow, or unused, farmers discovered that a wheat-sorghum-fallow cycle worked better. They also enacted other measures, such as the planting of trees to serve as windbreaks around croplands.

The Striking Landscape of Erosion

Among the by-products of erosion are some of the most dramatic landscapes in the world, many of which are to be found in the United States. A particularly striking example appears in Colorado, where the Arkansas River carved out the Royal Gorge. Though it is not nearly as deep as the Grand Canyon, this one has something the more famous gorge does not: a bridge. Motorists with the stomach for it can cross a span 1,053 ft. (0.32 km) above the river, one of the most harrowing drives in America.

Another, perhaps equally taxing, drive is that down California 1, a gorgeous scenic highway whose most dramatic stretches lie between Carmel and San Simeon. Drivers headed south find themselves pressed up against the edge of the cliffs, such that the slightest deviation from the narrow road would send an automobile and its passengers plummeting to the rocks many hundreds of feet below. These magnificent, terrifying landforms are yet another product of erosion, in this case, the result of the pounding Pacific waves.

Also striking is the topography produced by the erosion of material left over from a volcanic eruption. As discussed in the Mountains essay, Devils Tower National Monument in Wyoming is

KEY TERMS

CREEP: A form of mass wasting involving the slow downward movement of regolith as a result of gravitational force.

DELTA: A region of sediment formed when a river enters a larger body of water, at which point the reduction in velocity on the part of the river current leads to the widespread deposition of sediment.

DEPOSITION: The process whereby sediment is laid down on the Earth's surface.

EROSION: The movement of soil and rock due to forces produced by water, wind, glaciers, gravity, and other influences. In most cases, a fluid medium, such as air or water, is involved.

FLOW: A form of mass wasting in which a body of material that is not uniform moves rapidly downslope.

GEOMORPHOLOGY: An area of physical geology concerned with the study of landforms, with the forces and processes that have shaped them, and with the description and classification of various physical features on Earth.

GLACIER: A large, typically moving mass of ice either on or adjacent to a land surface.

LANDFORM: A notable topographical feature, such as a mountain, plateau, or valley.

MASS WASTING: The transfer of earth material, by processes that include creep, slump, slide, flow, and fall, down slopes. Also known as *mass movement.*

MORPHOLOGY: Structure or form or the study thereof.

REGOLITH: A general term describing a layer of weathered material that rests atop bedrock.

SEDIMENT: Material deposited at or near Earth's surface from a number of sources, most notably preexisting rock.

SLIDE: A variety of mass wasting in which material moves downhill in a fairly coherent mass (i.e., more or less in a section or group) along a flat or planar surface.

SLUMP: A form of mass wasting that occurs when a mass of regolith slides over or creates a concave surface (one shaped like the inside of a bowl).

TOPOGRAPHY: The configuration of Earth's surface, including its relief as well as the position of physical features.

WEATHERING: The breakdown of rocks and minerals at or near the surface of Earth due to physical, chemical, or biological processes.

the remains of an extinct volcano whose outer surface long ago eroded, leaving just the hard lava of the volcanic "neck." Erosion of lava also can produce mesas. Lava that has settled in a river valley may be harder than the rocks of the valley walls, such that the river eventually erodes the rocks, leaving only the lava platform. What was once the floor of the valley thus becomes the top of a mesa.

CONTROLLING EROSION

The force that shapes valleys and coastlines is certainly enough to destroy hill slopes, often with disastrous consequences for nearby residents. Such has been the case in California, where, during the 1990s, areas were dealt a powerful one-two punch of drought followed by rain. The drought killed off much of the vegetation that might have held the hillsides, and when rains came, they brought about mass wasting in the form of mudflows and landslides.

Over the surface of the planet, the average rate of erosion is about 1 in. (2.2 cm) in a thousand years. This is the *average*, however, meaning that in some places the rate is much, much higher, and in others it is greatly lower. The rate of erosion depends on several factors, including climate, the nature of the materials, the slope and angle of repose, and the role of plant and animal life in the local environment.

Whereas many types of plants help prevent erosion, the wrong types of planting can be detrimental. The dangers of improper land usage for crops and livestock are illustrated by the Dust Bowl experience, which highlights the fact that the organism most responsible for erosion is humanity itself. On the other hand, people also can protect against erosion by planting vegetation that holds the soil, by carefully managing and controlling land usage, and by lessening slope angle in places where gravity tends to erode the soil.

WHERE TO LEARN MORE

Cherrington, Mark. *Degradation of the Land.* New York: Chelsea House, 1991.

"Coastal and Nearshore Erosion." United States Geological Survey (USGS) (Web site). <http://walrus.wr.usgs.gov/hazards/erosion.html>.

Dean, Cornelia. *Against the Tide: The Battle for America's Beaches.* New York: Columbia University Press, 1999.

Hecht, Jeff. *Shifting Shores: Rising Seas, Retreating Coastlines.* New York: Scribners, 1990.

Middleton, Nick. *Atlas of Environmental Issues.* Illus. Steve Weston and John Downes. New York: Facts on File, 1989.

Protecting Your Property from Erosion (Web site). <http://www.abag.ca.gov/bayarea/enviro/erosion/erosion.html>.

Rybolt, Thomas R., and Robert C. Mebane. *Environmental Experiments About Land.* Hillside, NJ: Enslow Publishers, 1993.

"Soil Erosion on Farmland." New Zealand Ministry of Agriculture and Forestry (Web site). <http://www.maf.govt.nz/MAFnet/publications/erosion-risks/httoc.htm>.

Weathering and Erosion (Web site). <http://vishnu.glg.nau.edu/people/jhw/GLG101/Weathering.html>.

Wind Erosion Research Unit. United States Department of Agriculture/Kansas State University (Web site). <http://www.weru.ksu.edu/>.

MASS WASTING

CONCEPT

The term mass wasting (sometimes called mass movement) encompasses a broad array of processes whereby earth material is transported down a slope by the force of gravity. It is related closely to weathering, which is the breakdown of minerals or rocks at or near Earth's surface through physical, chemical, or biological processes, and to erosion, the transport of material through a variety of agents, most of them flowing media, such as air or water. Varieties of mass wasting are classified according to the speed and force of the process, from extremely slow creep to very rapid, dramatic slide or fall. Examples of rapid mass wasting include landslides and avalanches, which can be the cause of widespread death and destruction when they occur in populated areas.

HOW IT WORKS

Moving Earth and Rocks

In discussing mass wasting, the area of principal concern is Earth's surface rather than its interior. Thus, mass wasting is related most closely to the realm of geomorphology, a branch of physical geology concerned with the study of landforms, with the forces and processes that have shaped them, and with the description and classification of various physical features on Earth. Though plate tectonics (which involves the movement of giant plates beneath the earth's surface) can influence mass wasting, plate tectonics entails interior processes that humans usually witness only indirectly, by seeing their effects. Mass wasting, on the other hand, often can be observed directly, particularly in its more rapid forms, such as rock fall.

There are three general processes whereby a piece of earth material can be moved from a high outcropping to the sea: weathering, mass wasting, or erosion. If mechanical, biological, or chemical processes act on the material, dislodging it from a larger sample of material (e.g., separating a rock from a boulder), it is an example of weathering, which is discussed later in this essay. Supposing that a rock has been broken apart by weathering, it may be moved further by mass-wasting processes, such as creep or fall. Pieces of rock swept away by a river in a valley below the outcropping and small bits of rock worn away by high winds are examples of erosion. Erosion and weathering are examined in separate essays within this book.

As for the relationships between erosion, weathering, and mass wasting, the lines are not clearly drawn. Some authors treat weathering and mass wasting as varieties of erosion, and some apply a strict definition of erosion as resulting only from flowing media. (In the physical sciences, fluid means anything that flows, not just liquids.) Weathering, mass wasting, and erosion also can be viewed as stages in a process, as described in the preceding paragraph. This broad array of approaches, while perhaps confusing, only serves to illustrate the fact that the earth sciences are relatively young compared with such ancient disciplines as astronomy and biology. Not all definitions in the earth sciences are, as it were, "written in stone."

Weathering

A mineral is a substance that occurs naturally, is usually inorganic, and typically has a crystalline structure. The term *organic* does not necessarily mean "living" rather, it refers to all carbon-containing compounds other than oxides, such as carbon dioxide, and carbonates, which are often found in Earth's rocks. A crystalline solid is one in which the constituent parts have a simple and definite geometric arrangement repeated in all directions.

Rocks, scientifically speaking, are simply aggregates or combinations of minerals or organic material or both, and weathering is the process whereby rocks and minerals are broken down into simpler materials. Weathering is the mechanism through which soil is formed, and therefore it is a geomorphologic process essential to the sustenance of life on Earth. There are three varieties of weathering: physical or mechanical, chemical, and biological.

THE THREE TYPES OF WEATHERING. Physical or mechanical weathering involves such factors as gravity, friction, temperature, and moisture. Gravity, for instance, may cause a rock to drop from a height, such that it falls to the ground and breaks into pieces. If wind-borne sand blows constantly across a rock surface, the friction will have the effect of sandpaper, producing mechanical weathering. In addition, changes in temperature and moisture will cause expansion and contraction of materials, bringing about sometimes dramatic changes in their physical structure.

Chemical weathering not only is a separate variety of weathering but also is regarded as a second stage, one that follows physical weathering. Whereas physical changes are typically external, chemical changes affect the molecular structure of a substance, bringing about a rearrangement in the ways that atoms are bonded. Important processes that play a part in chemical weathering include acid reactions, hydrolysis (a reaction with water that results in the separation of a compound to form a new substance or substances), and oxidation. The latter can be defined as any chemical reaction in which oxygen is added to or hydrogen is removed from a substance.

An example of biological weathering occurs when a plant grows from a crevice in a rock. As the plant grows, it gradually forces the sides of the crevice apart even further, and it ultimately may tear the rock apart. Among the most notable agents of biological weathering are algae and fungi, which may be combined in a mutually beneficial organism called a lichen. (Reindeer moss is an example of a lichen.) Through a combination of physical and chemical processes, organisms ranging from lichen to large animals can wear away rock gradually.

Properties of Unconsolidated Material

Regolith is a general term that describes a layer of weathered material that rests atop bedrock. It is unconsolidated, meaning that it is in pieces, like gravel, though much less uniform in size. Sand and soil, including soil mixed with loose rocks, are examples of regolith.

Every variety of unconsolidated material has its own angle of repose, or the maximum angle at which it can remain standing. Everyone who has ever attempted to build a sand castle at the beach has experienced angle of repose firsthand, perhaps without knowing it. Imagine that you are trying to build a sand castle with a steep roof. Dry sand would not be good for this purpose, because it is loose and has a tendency to flow easily. Much better would be moist sand, which can be shaped into a sharper angle, meaning that it has a higher angle of repose.

A certain amount of water gives sand surface tension, the same property that causes water to bead up on a table rather than lying flat. If too much water is added to the sand, however, the sand becomes saturated and will flow, a process called lateral spreading. Thus, to a point, the addition of water increases the angle of repose for sand, which is only about 34° when the sand is dry. (This is the angle of repose for sand in an hourglass.) On the other hand, piles of rocks may have an angle of repose as high as 45°. In practice, most aggregates of materials in nature have slopes less than their angle of repose, owing to the influence of wind and other erosive forces.

Types of Mass Wasting

As noted earlier, there is some disagreement among writers in the geologic sciences regarding the types of mass wasting. Indeed, even the term mass wasting is not universal, since some writers refer to it as mass movement. Others do not even treat the subject as a category unto itself, preferring instead to address related concepts, such as weathering and erosion, as well as instances of mass wasting, such as avalanches and landslides.

AN AVALANCHE ON MOUNT MCKINLEY IN ALASKA. *(© W. Bacon/Photo Researchers. Reproduced by permission.)*

For this reason, the classification of mass-wasting processes presented here is by no means universal and instead represents a composite of several schools of thought. Generally speaking, geologists and geomorphologists classify processes of mass wasting according to the rapidity with which they occur. Most sources recognize at least three types of mass wasting: flow, slide, and fall. Some sources include slump among the categories of relatively rapid mass-wasting process, as opposed to the slower, less dramatic (but ultimately more important) process known as creep. Some writers classify uplift and subsidence with mass wasting; however, in this book, uplift and subsidence are treated separately, in the Geomorphology essay.

REAL-LIFE APPLICATIONS

CREEP

Creep is the slow downward movement of regolith as a result of gravitational force. Before the initiation of the creeping process, the regolith is in what physicists call a condition of unstable equilibrium: it remains in place, yet a relatively small disturbance would be enough to dislodge

A mudflow caused by heavy winter rains brings down the hillside under homes in Millbrae, California. *(AP/Wide World Photos. Reproduced by permission.)*

it. Though it is slow, creep can produce some of the most dramatic results over time. It can curve tree trunks at the base, break or overturn retaining walls, and cause objects from fence posts to utility poles to tombstones to be overturned.

Changes in temperature or moisture are among the leading factors that result in the disturbance of regolith. A change in either can cause material to expand or contract, and freezing or thawing may be enough to shake regolith from its position of unstable equilibrium. In fact, some geomorphologists cite a distinct mass-wasting process, known as solifluction, that occurs in the active layer of permafrost, which thaws in the summertime. Water also can provide lubrication or additional weight that assists the material in moving. One of the only causes of creep not associated with changes in temperature or moisture is the burrowing of small animals.

Slump and Slide

Slump occurs when a mass of regolith slides over or creates a concave surface (one shaped like the inside of a bowl). The result is the formation of a small, crescent-shaped cliff, known as a scarp, at the upper end—rather like the crest of a wave. Soil flow takes place at the bottom end of the slump. One is likely to see slumps in any place where forces, whether man-made or natural, have graded material to a slope too steep for its angle of repose. This may happen along an interstate highway, where a road crew has cut the slope too sharply, or on a riverbank, where natural erosion has done its work.

Often, slump is classified as a variety of slide, in which material moves downhill in a fairly coherent mass (i.e., more or less in a section or group) along a flat or planar surface. These movements sometimes are called rock slides, debris slides, or, in common parlance, landslides. Among the most destructive types of mass wasting, they may be set in motion by earthquakes, which are caused by plate tectonic processes, or by hydrologic agents (i.e., excessive rain or melting snow and ice).

Flow

When a less uniform, or more chaotic, mass of material moves rapidly downslope, it is called flow. Flow is divided into categories, depending on the amounts of water involved: granular flow (0-20% water) and slurry flow (20-40% water). Creep and solifluction often are classified as very slow forms of granular and slurry flow, respectively. In order of relative speed, these categories are as follows:

Granular Flow (0-20% Water)

- Slowest: Creep
- Slower: Earth flow
- Faster: Grain flow
- Fastest: Debris avalanche

Slurry Flow (20-40% Water)

- Slow: Solifluction
- Medium: Debris flow
- Fast: Mudflow

Earth flow moves at a rate anywhere from 3.3 ft. (1 m) per year to 330 ft. (100 m) per hour. Grain flow can be nearly 60 mi. (100 km) per hour, and debris avalanche may achieve speeds of 250 mi. (400 km) per hour, making it extremely dangerous. Among types of slurry flow, debris flow is roughly analogous to earth flow, falling into a range from about 4 in. (10 cm) per year to 0.6 mi. (1 km) per hour. Mudflow is slightly faster than grain flow. If the water content is more than 40%, a slurry flow is considered a stream.

Earth flows involve fine-grained materials, such as clay or silt, and typically occur in humid areas after heavy rains or the melting of snow. Debris flows usually result from heavy rains as well and may start with slumps before flowing downhill, forming lobes with a surface broken by ridges and furrows. Grain flows can be caused by a small disturbance, which forces the dry, unconsolidated material rapidly downslope. Debris avalanches are commonly the result of earthquakes or volcanic eruptions.

Seismic disturbances or volcanic activity may cause the collapse of a mountain slope, sending debris avalanches moving swiftly even along the gentler slopes of the mountainside. Likewise, mudflows may be the result of volcanic activity, in which case they are known as lahars. In some situations, the material in a lahar is extremely hot. Mudflows tend to be highly fluid mixtures of sediment (material deposited at or near Earth's surface from a number of sources, most notably preexisting rock) and water and typically flow along valley floors.

Fall

Most other forms of mass wasting entail movement along slopes that are considerably less than 90°, whereas fall takes place at angles almost perpendicular to the ground. Anyone who has driven through a wide mountain area, with steep cliffs on either side, has seen signs that say "Watch for Falling Rock." These warnings, which appear regularly on the drive through the Rockies in Colorado or on highways across the Blue Ridge and Great Smoky mountains in the southern United States, indicate the threat of rock fall.

The mechanism behind rock fall is simple enough. When a rock at the top of a slope is in unstable equilibrium, it can be dislodged such that it either falls directly downward or bounces and rolls. Usually, the bottom of the slope or cliff contains accumulated talus, or fallen rock material. Freezing and thawing as well as the growth of plant roots may cause fall. The latter is not limited to rock fall: debris fall, which is closely related, includes soil, vegetation, and regolith as well as rocks.

Mass Wasting and Natural Disasters

Among the most dramatic and well-known varieties of mass wasting are avalanches, a variety of flow, and landslides, which (as their name suggests) are a type of slide. These can result, and have resulted, in enormous loss of life and property. Some notable modern occurrences of mass wasting, and the type of movement involved, are listed below. With each incident, the approximate number of fatalities is shown in parentheses.

- *China, 1920:* Landslide caused by an earthquake (200,000)
- *Peru, 1970:* Debris avalanche related to an earthquake (70,000)
- *Colombia, 1985:* Mudflow related to a volcanic eruption (23,000)
- *Soviet Union, 1949:* Landslide caused by an earthquake (12,000–20,000)
- *Italy and Austria, 1916:* Landslide (10,000)
- *Peru, 1962:* Landslide (4,000–5,000)
- *Italy, 1963:* Landslide (2,000)
- *Japan, 1945:* Landslide caused by a flood (1,200)
- *Ecuador, 1987:* Landslide related to an earthquake (1,000)
- *Austria, 1954:* Landslide (200)

The Role of Plate Tectonics

Note how many times an instance of mass wasting was either caused by or "related to" (meaning that geologists could not establish a full causal relationship) volcanic or seismic activity. Both, in

KEY TERMS

ANGLE OF REPOSE: The maximum slope at which a relatively large sample of unconsolidated material can remain standing. Often, the addition of water increases the angle of repose, up to the point at which the material becomes saturated.

AVALANCHE: See *flow.*

CREEP: A form of mass wasting involving the slow downward movement of regolith as a result of gravitational force.

EROSION: The movement of soil and rock due to forces produced by water, wind, glaciers, gravity, and other influences. In most cases, a fluid medium, such as air or water, plays a part.

FALL: A form of mass wasting in which rock or debris moves downward along extremely steep angles.

FLOW: A form of mass wasting in which a body of material that is not uniform moves rapidly downslope. Flow is divided into categories, depending on the amounts of water involved: granular flow (0-20% water) and slurry flow (20-40% water). An avalanche is an example of flow and may involve either rock (granular) or snow (slurry).

FLUID: In the physical sciences, the term fluid refers to any substance that flows and therefore has no definite shape—that is, both liquids and gases. In the earth sciences, occasionally substances that appear to be solid, for example, ice in glaciers, are, in fact, flowing slowly.

GEOMORPHOLOGY: An area of physical geology concerned with the study of landforms, with the forces and processes that have shaped them, and with the description and classification of various physical features on Earth.

LANDSLIDE: See *slide.*

LITHOSPHERE: The upper layer of Earth's interior, including the crust and the brittle portion at the top of the mantle.

turn, are the result of plate movement in most instances, and thus it is not surprising that several of the locales noted here are either at plate margins or in mountainous regions where plate tectonic and other processes are at work. (For more on this subject, see the entries Plate Tectonics and Mountains.)

To set mass wasting into motion, it is necessary to have a steep slope and some type of force to remove material from its position of unstable equilibrium. Plate tectonic processes provide both. Not only does an earthquake, for instance, jar rocks loose from the upper portion of a slope, but the movement of plates also helps create steep slopes, for example, the collision of the Indo-Australian and Eurasian belts that produced the Himalayas.

Some of the most vigorous plate tectonic activity occurs underwater, and, likewise, there are remarkable manifestations of mass wasting beneath the seas. Off Moss Landing, a research facility that serves a consortium of state universities in northern California, is an underwater canyon more than 0.6 mi. (1 km) deep. At one time, Monterey Canyon was thought to be the result of erosion by a river flowing into the ocean; however, today it is believed to be the result of underwater mass wasting.

Detecting and Preventing Mass Wasting

The dramatic instances of mass wasting discussed here hardly require any effort at detection. Their effect is obvious and, to those unfortunate enough

KEY TERMS

MASS WASTING: The transfer of earth material down slopes by processes that include creep, slump, slide, flow, and fall. Also known as *mass movement.*

PLATE MARGINS: Boundaries between plates.

PLATE TECTONICS: The name both of a theory and of a specialization of tectonics. As an area of study, plate tectonics deals with the large features of the lithosphere and the forces that shape them. As a theory, it explains the processes that have shaped Earth in terms of plates and their movement.

PLATES: Large, movable segments of the lithosphere.

REGOLITH: A general term describing a layer of weathered material that rests atop bedrock.

SEDIMENT: Material deposited at or near Earth's surface from a number of sources, most notably preexisting rock.

SLIDE: A variety of mass wasting in which material moves downhill in a fairly coherent mass (i.e., more or less in a section or group) along a flat or planar surface.

SLUMP: A form of mass wasting that occurs when a mass of regolith slides over or creates a concave surface (one shaped like the inside of a bowl).

SURFACE TENSION: An attractive force exerted by molecules in the interior of a liquid on molecules at the exterior. This force draws the material inward such that it occupies less than its maximum horizontal area. The surface tension of water is high, causing it to bead on most surfaces.

UNSTABLE EQUILIBRIUM: A situation in which an object remains in place, yet a relatively small disturbance would be enough to dislodge it.

WEATHERING: The breakdown of rocks and minerals at or near the surface of Earth due to physical, chemical, or biological processes.

to be nearby, inescapable. Other types of mass wasting occur so slowly that they do not invite immediate detection. This can be unfortunate, because in some cases slow mass wasting is a harbinger of much more rapid movements to follow.

A dwelling atop a hill is subject to enormous gravitational force, and the more massive the dwelling, the greater the pull of gravity. (Weight is, after all, nothing but gravitational force.) If a homeowner adds a swimming pool or other items that contribute to the weight of the dwelling, it only increases the chances that it may experience mass wasting. Heavy rains can bring so much water that it saturates the soil, reducing its surface tension and causing it to slide—as occurred, for instance, in the area around Malibu, California, during the late 1990s.

The California mud slides and landslides are a dramatic example of mass wasting, but more often than not mass wasting takes the form of creep, which is detectable only over a matter of years. When creep occurs, the upper layer of soil moves, while the layer below remains stationary. One way to keep the upper layer in place is to plant vegetation that will put down roots deep enough to hold the soil.

This may create unintended consequences. During the 1930s, New Deal officials imported kudzu plants from China, intending to protect the hillsides of the American South from creep and erosion. The kudzu protected the slopes, but as it turned out, this voracious plant had a tendency to creep as well. Before communities began taking steps to eradicate it, or at least push it back, in the 1970s, kudzu seemingly threatened to cover the entire southern United States.

To prevent some of the more dramatic varieties of mass wasting, such as landslides in a residential area, a homeowner or group of homeowners may commission an engineer's study. The engineer can test the material of the slope, measure the stresses acting on it, and perform other calculations to predict the likelihood that a slope will succumb to a given amount of force. For this reason, zoning laws in areas with steep slopes are typically strict. These laws are geared toward preventing homeowners and builders from erecting structures likely to create a threat of mass wasting in a period of heavy rains.

WHERE TO LEARN MORE

Abbott, Patrick. *Natural Disasters.* Dubuque, IA: William C. Brown Publishers, 1996.

Allen, Missy, and Michel Peissel. *Dangerous Natural Phenomena.* New York: Chelsea House, 1993.

Armstrong, Betsy R., and Knox Williams. *The Avalanche Book.* Golden, CO: Fulcrum, 1986.

Goodwin, Peter. *Landslides, Slumps, and Creeps.* New York: Franklin Watts, 1997.

Gore, Pamela. "Mass Wasting" (Web site). <http://www.gpc.peachnet.edu/~pgore/geology/geo101/masswasting.html>.

Mass Wasting (Web site). <http://www.es.ucsc.edu/~jsr/EART10/Lectures/HTML/lecture.07.html>.

"Mass Wasting Features of North Dakota." North Dakota State University (Web site). <http://www.ndsu.nodak.edu/nd_geology/nd_mwast/index_mw.htm>.

Murck, Barbara Winifred, Brian J. Skinner, and Stephen C. Porter. *Dangerous Earth: An Introduction to Geologic Hazards.* New York: John Wiley, 1996.

Nelson, Stephen A. *Mass-Wasting and Mass-Wasting Processes. Tulane University* (Web site). <http://www.tulane.edu/~sanelson/geol204/masswastproc.htm>.

Weathering and Mass Wasting Learning Module (Web site). <http://home.aol.com/rhaberlin/mwmod.htm>.

SEDIMENTOLOGY AND SOIL SCIENCE

SEDIMENT AND SEDIMENTATION

CONCEPT

The materials that make up Earth are each products of complex cycles and interactions, as a study of sediment and sedimentation shows. Sediment is unconsolidated material deposited at or near Earth's surface from a number of sources, most notably preexisting rock. There are three kinds of sediment: chemical, organic, and rock, or clastic sediment. Weathering removes this material from its source, while erosion and mass wasting push it along to a place where it is deposited. After deposition, the material may become a permanent part of its environment, or it may continue to undergo a series of cycles in which it experiences ongoing transformation.

HOW IT WORKS

Transporting Sediment

There are three types of sediment: rocks, or clastic sediment; mineral deposits, or chemical sediment; and organic sediment, composed primarily of organic material. (In this context, the term *organic* refers to formerly living things and parts or products of living things; however, as discussed in Minerals, the term actually has a much broader meaning.) There are also three general processes involved in the transport of sediment from higher altitudes to lower ones, where they eventually are deposited: weathering, mass wasting, and erosion.

The lines between these three processes are not always clearly drawn, but, in general, the following guidelines apply. When various processes act on the material, causing it to be dislodged from a larger sample (for example, separating a rock from a boulder), this is an example of weathering. Assuming that a rock has been broken apart by weathering, it may be moved farther by mass-wasting processes, such as creep or fall, for which gravity is the driving factor. If the pieces of rock are swept away by a river, high winds, or a glacier (all of which are flowing media), this is an example of erosion.

WEATHERING. Weathering is divided further into three different types: physical, chemical, and biological. Physical or mechanical weathering takes place as a result of such factors as gravity, friction, temperature, and moisture. Gravity may cause a rock to roll down a hillside, breaking to pieces at the bottom; friction from particles of matter borne by the wind may wear down a rock surface; and changes in temperature and moisture can cause expansion and contraction of materials.

Whereas physical weathering attacks the rock as a whole, chemical weathering involves the breakdown of the minerals or organic materials that make up the rock. Chemical breakdown may lead to the dissolution of the materials in the rock, which then are washed away by water or wind or both, or it may be merely a matter of breaking the materials down into simpler compounds.

Biological weathering is not so much a third type of weathering as it is a manifestation of chemical and physical breakdown caused by living organisms. Suppose, for instance, that a plant grows within a crack in a rock. Over time, the plant will influence physical weathering through its moisture and the steady force of its growth pushing at the walls of the fissure in which it is rooted. At the same time, its specific chemical

SEDIMENT CAN BE TRANSPORTED BY WEATHERING, EROSION, AND MASS WASTING. HERE A WINTER STORM HAS CAUSED COASTAL EROSION OF A BEACH. *(© Rafael Macia/Photo Researchers. Reproduced by permission.)*

properties likely will induce decomposition of the rock.

MASS WASTING. Mass wasting, sometimes known as mass movement, comprises a number of types of movement of earth material, all of them driven by gravity. Creep is the slow downward drift of regolith (unconsolidated material produced by weathering), while slump occurs when a mass of regolith slides over, or creates, a concave surface—that is, one shaped like the inside of a bowl. Slump sometimes is classified as a variety of slide, in which material moves downhill in a fairly coherent mass along a flat or planar surface. Such movements, sometimes called rock slides, debris slides, or landslides, are among the most destructive types of mass wasting.

When a less uniform, or more chaotic, mass of material moves rapidly down a slope, it is called flow. Flow is divided into categories, depending on the specific amounts of water: granular flows (0–20% water) and slurry flows (20–40% water), the fastest varieties of which are debris avalanche and mudflow, respectively. Mudflows can be more than 60 mi. (100 km) per hour, while debris avalanches may achieve speeds of 250 mi. (400 km) per hour.

Even these high-speed varieties of mass wasting entail movement along slopes that are considerably less than 90°, whereas a final variety of mass wasting, that is, fall, takes place at angles almost perpendicular to the ground. Typically the bottom of a slope or cliff contains accumulated talus, or fallen rock material. Nor is fall limited to rock fall: debris fall, which is closely related, includes soil, vegetation, and regolith as well as rocks. (For more on these subjects, see Mass Wasting.)

EROSION

Erosion typically is caused either by gravity (in which case it is generally known as mass wasting, discussed earlier) or by flowing media, such as water, wind, and even ice in glaciers. It removes sediments in one of three ways: by the direct impact of the agent (i.e., the flowing media that is discussed in the following sections); by abrasion, another physical process; or by corrosion, a chemical process.

In the case of direct impact, the wind, water, or ice removes sediment, which may or may not be loose when it is hit. On the other hand, abrasion involves the impact of solid earth materials carried by the flowing agent rather than the impact of the flowing agent itself. For example, sand borne by the wind, as discussed later, or pebbles carried by water may cause abrasion.

Corrosion is chemical and is primarily a factor only in water-driven, as opposed to wind-driven or ice-driven, erosion. Streams slowly dissolve rock, removing minerals that are carried downstream by the water.

WATER. Of the fluid substances driving erosion, liquid water is perhaps the one most readily associated in most people's minds with erosion. In addition to the erosive power of waves on the seashore, there is the force exerted by running water in creeks, streams, and rivers. As a river moves, pushing along sediment eroded from the streambeds or riverbeds, it carves out deep chasms in the bedrock beneath.

Moving bodies of water continually reshape the land, carrying soil and debris down slopes, from the source of the river to its mouth, or delta. A delta is a region formed when a river enters a larger body of water, at which point the reduction in velocity on the part of the river current leads to the widespread deposition of sediment. It is so named because its triangular shape resembles that of the Greek letter delta, Δ.

Water at the bottom of a large body, such as a pond or lake, also exerts erosive power; then there is the influence of falling rain. Assuming that the ground is not protected by vegetation, raindrops can loosen particles of soil, sending them scattering in all directions. A rain that is heavy enough may dislodge whole layers of topsoil and send them rushing away in a swiftly moving current. The land left behind may be rutted and scarred, much of its best soil lost for good.

ICE. Ice, of course, is simply another form of water, but since it is solid, its physical properties are quite different. It is a solid rather than a fluid, such as liquid water or air (the physical sciences treat gases and liquids collectively as "fluids"), yet owing to the enormous volume of ice in glaciers, these great masses are capable of flowing. Glaciers do not flow in the same way as a fluid does; instead, they are moved by gravity, and like giant bulldozers made of ice, they plow through rock, soil, and plants.

Under certain conditions a glacier may have a layer of melted water surrounding it, which greatly enhances its mobility. Even without such lubricant, however, these immense rivers of ice move steadily forward, gouging out pieces of bedrock from mountain slopes, fashioning deep valleys, removing sediment from some regions and adding it to others. In unglaciated areas, or places that have never experienced any glacial activity, sediment is formed by the weathering and decomposition of rock. On the other hand, formerly glaciated areas are distinguished by layers of till, or glacial sediment, from 200 to 1,200 ft. (61–366 m) thick.

WIND. The processes of wind erosion sometimes are called *eolian* processes, after Aeolus, the Greek god of the winds. Eolian erosion is in some ways less forceful than the erosive influence of water. Water, after all, can lift heavier and larger particles than can the winds. Wind, however, has a much greater frictional component in certain situations. This is particularly true when the wind carries sand, every grain of which is like a cutting tool.

Wind erosion, in fact, is most pronounced in precisely those places where sand abounds, in deserts and other areas that lack effective ground cover in the form of solidly rooted, prevalent vegetation. In some desert regions the bases of rocks or cliffs have been sandblasted, leaving a mushroom-shaped formation owing to the fact that the wind could not lift the fine grains of sand very high.

Sediment Load

Eroded particles become part of what is called the sediment load transported by the fluid medium. Sediment load falls into three categories: dissolved load, suspended load, and bed load. The amount of each type of load that a fluid medium is capable of carrying depends on the density of the fluid medium itself: in other words, wind can carry the least of each and ice the most.

The wind does not carry any dissolved load, since solid particles (unlike gases) cannot be dissolved in air. Ice or water, on the other hand, is able to dissolve materials, which become invisible within them. Typically, about 90% of the dissolved load in a river is accounted for by five different ions, or atoms that carry a net electric charge: the anions (negative ions) chloride, sulfate, and bicarbonate and the cations (positive ions) of sodium and calcium.

Suspended load is sediment that is suspended, or floating, in the erosive medium. In this instance, wind is just as capable as water or ice of suspending particles of the sediment load, which are likely to color the medium that carries them. Hence, water or wind carrying suspended particles is usually murky. The thicker the medium, the larger the particles it is capable of suspending. In other words, ice can suspend extremely large pieces of sediment, whereas water can suspend much more modest ones. Wind can suspend only tiny particles.

Then there is bed load, large sediment that never becomes suspended but rather is almost always in contact with the substrate or bottom, whether "the bottom" is a streambed or the ground itself. Instead of being lifted up by the medium, bed load is nudged along, rolling, skipping, and sliding as it makes its way over the substrate. Once again, the density of the medium itself has a direct relationship to the size of the bed load it is capable of carrying. Wind rarely transports bed load thicker than fine sand, and water usually moves only pebbles, though under flood conditions it can transport boulders. As with suspended load, glaciers can transport virtually any size of bed load.

Sediment Sizes and Shapes

Geologists and sedimentologists use certain terms to indicate sizes of the individual particles in sediment. Many of these terms are familiar to us from daily life, but whereas people typically use them in a rather vague way, within the realm of sedimentology they have very specific meanings. Listed below are the various sizes of rock, each with measurements or measurement ranges for the rock's diameter:

- Clay: Smaller than 0.00015 in. (0.004 mm)
- Silt: 0.00015 in. (0.004 mm) to 0.0025 in. (0.0625 mm)
- Sand: 0.0025 in. (0.0625 mm) to 0.08 in. (2 mm)
- Pebble: 0.08 in. (2 mm) to 2.5 in. (64 mm)
- Cobble: 2.5 in. (64 mm) to 10 in. (256 mm)
- Boulder: Larger than 10 in. (256 mm).

This listing is known as the Udden–Wentworth scale, which was developed in 1898 by J.A. Udden (1859–1932), an American sedimentary petrologist (a scientist who studies rocks). In 1922 the British sedimentary petrologist C. K. Wentworth) expanded Udden's scale, adapting the definitions of various particle sizes to fit more closely with the actual usage and experience of researchers in the field. The scale uses modifiers to pinpoint the relative sizes of particles. In ascending order of size, these sizes are very fine, fine, medium, coarse, and very coarse.

Real-Life Applications

Sediments and Dust Bowls

Sediment makes possible the formation of soil, which of course is essential for growing crops. Therefore it is a serious matter indeed when wind and other forces of erosion remove sediment, creating dust-bowl conditions. The term "Dust Bowl," with capital letters, refers to the situation that struck the United States Great Plains states during the 1930s, devastating farms and leaving thousands of families without home or livelihood. (See Erosion for much more about the Dust Bowl.)

During the late 1990s, some environmentalists became concerned that farming practices in the western United States were eroding sediment, putting in place the possibility of a return to the conditions that created the Dust Bowl. However, in August 1999, the respected journal *Science* reported studies showing that sediment in farmlands was not eroding at anything like the rate that had been feared. Soil scientist Stanley Trimble at the University of California, Los Angeles, studied Coon Creek, Wisconsin, and its tributaries, a watershed for which 140 years' worth of erosion data were available. As Trimble discovered, the rate of sediment erosion in the area had dramatically decreased since the 1930s, and was now at 6% of the rate during the Dust Bowl years.

Some studies from the 1970s onward had indicated that farming techniques, designed to improve the crop output from the soil, had created a situation in which sediment was being washed away at alarming rates. However, if such sediment removal were actually taking place, there would have to be some evidence—if nothing more, the sediment that had been washed away would have had to go somewhere. Instead, as Trimble reported, "We found that much of the sediment in Coon Creek doesn't move very far, and that it moves in complex ways." The sediment, as he went on to explain, was moving within the Coon Creek basin, but the amount that actually made it to the Mississippi River (which could be counted as true erosion, since it was removing sediment from the area) had stayed essentially the same for the past 140 years.

SEDIMENTARY STRUCTURES REMAINING IN A DRIED RIVER BED. CLAY SOILS CRACK AS THEY LOSE MOISTURE AND CONTRACT WHEN TRAPPED WATER EVAPORATES AS THE RESULT OF DROUGHT. *(© B. Edmaier/Photo Researchers. Reproduced by permission.)*

DEPOSITION AND DEPOSITIONAL ENVIRONMENTS

Eventually everything in motion—including sediment—comes to rest somewhere. A piece of sediment traveling on a stream of water may stop hundreds of times, but there comes a point when it comes to a complete stop. This process of coming to rest is known as deposition, which may be of two types, mechanical or chemical. The first of these affects clastic and organic sediment, while the second applies (fittingly enough) to chemical sediment.

In mechanical deposition, particles are deposited in order of their relative size, the largest pieces of bed load coming to a stop first. These large pieces are followed by medium-size pieces and so on until both bed load and suspended load have been deposited. If the sediment has come to a full stop, as, for instance, in a stagnant pool of water, even the finest clay suspended in the water eventually will be deposited as well.

Unlike mechanical deposition, chemical deposition is not the result of a decrease in the velocity of the flow; rather, it comes about as a result of chemical precipitation, when a solid particle crystallizes from a fluid medium. This often happens in a saltwater environment, where waters may become overloaded with salt and other minerals. In such a situation, the water is unable to maintain the minerals in a dissolved state (i.e., in solution) and precipitates part of its content in the form of solids.

DEPOSITIONAL ENVIRONMENTS. The matter of sediment deposition in water is particularly important where reservoirs are concerned, since in that case the water is to be used for drinking, cooking, bathing, and other purposes by humans. One of the biggest problems for the maintenance of clean reservoirs is the transport of sediment from agricultural areas, in which the soil is likely to contain pesticides and other chemicals, including the phosphorus found in fertilizer. A number of factors, including precipitation, topography, and land use, affect the rate at which sediment is deposited in reservoirs.

The area in which sediment is deposited is known as its depositional environment, of which there are three basic varieties: terrestrial, marginal marine, and marine. These are, respectively, environments on land (and in landlocked waterways, such as creeks or lakes), along coasts, and in the open ocean. A depositional environment may be a large-scale one, known as a regional environment, or it may be a smaller subenvironment, of which there may be hundreds within a given regional environment.

SEDIMENTARY STRUCTURES. There are many characteristic physical formations, called sedimentary structures, that sediment forms after it has reached a particular depositional environment. These formations include bedding planes and beds, channels, cross-beds, ripples, and mud cracks. A bed is a layer, or stratum, of sediment, and bedding planes are surfaces that separate beds. The bedding plane indicates an interruption in the regular order of deposition. (These are concepts that also apply to the field of stratigraphy. For more on that subject, see the essay Stratigraphy.)

Channels are simply depressions in a bed that reflect the larger elongated depression made by a river as it flows along its course. Cross-beds are portions of sediment that are at an angle to the beds above and below them, as a result of the action of wind and water currents—for example, in a flowing stream. As for ripples, they are small sandbar-like protuberances that form perpendicular to the direction of water flow. At the beach, if you wade out into the water and look down at your feet, you are likely to see ripples perpendicular to the direction of the waves. Finally, mud cracks are the sedimentary structures that remain when water trapped in a muddy pool evaporates. The clay, formerly at the bottom of the pool, begins to lose its moisture, and as it does, it cracks.

The Impact of Sediment

It is estimated that the world's rivers carry as much as 24 million tons (21,772,800 metric tons) of sediment to the oceans each year. There is also the sediment carried by wind, glaciers, and gravity. Where is it all going? The answer depends on the type of sediment. Clastic and organic sediment may wind up in a depositional environment and experience compaction and cementation in the process of becoming sedimentary rock. (For more on sedimentary rock, see Rocks.)

On the other hand, clastic and organic particles may be buried, but before becoming lithified (turned to rock), they once again may be exposed to wind and other forces of nature, in which case they go through the entire cycle again: weathering, erosion, transport, deposition, and burial. This cycle may repeat many times before the sediment finally winds up in a permanent depositional environment. In the latter case, particles of clastic and organic sediment ultimately may become part of the soil, which is discussed elsewhere in this book (See Soil).

A chemical sediment also may become part of the soil, or it may take part in one or more biogeochemical cycles (also discussed elsewhere; see Biogeochemical Cycles). These chemicals may wind up as water in underground reservoirs, as ice at Earth's poles, as gases in the atmosphere, as elements or compounds in living organisms, or as parts of rocks. Indeed, all three types of sediment—clastic, chemical, and organic—are part of what is known as the rock cycle, whereby rocks experience endlessly repeating phases of destruction and renewal. (See Rocks for more details.)

Sedimentary Mineral Deposits

Among the most interesting aspects of sediment are the mineral deposits it contains—deposits that may, in the case of placer gold, be of significant value. A placer deposit is a concentration of heavy minerals left behind by the effect of gravity on moving particles, and since gold is the densest of all metals other than uranium (which

KEY TERMS

BED LOAD: Sediment that is capable of being transported by an erosive medium (wind, water, or air) but only under conditions in which it remains in nearly constant contact with the substrate or bottom (e.g., a streambed or the ground). Bed load, along with dissolved load and suspended load, is one of three types of sediment load.

COMPOUND: A substance made up of atoms of more than one element chemically bonded to one another.

CONSOLIDATION: A process whereby materials become compacted, or experience an increase in density. This takes place through a number of processes, including recrystallization and cementation.

DEPOSITION: The process whereby sediment is laid down on the Earth's surface.

DIAGENESIS: A term referring to all the changes experienced by a sediment sample under conditions of low temperature and low pressure following deposition.

DISSOLVED LOAD: Sediment load that is absorbed completely by the erosive medium (either water or ice) that carries it. Dissolved load is one of three types of sediment load, the others being suspended load and bed load.

EROSION: The movement of soil and rock due to forces produced by water, wind, glaciers, gravity, and other influences. In most cases, a fluid medium, such as air or water, is involved.

FLUID: In the physical sciences, the term fluid refers to any substance that flows and therefore has no definite shape—that is, both liquids and gases. Occasionally, substances that appear to be solid (for example, ice in glaciers), in fact, are flowing slowly; therefore, within the earth sciences, ice often is treated as another fluid medium.

ION: An atom or group of atoms that has lost or gained one or more electrons and thus has a net electric charge. Positively charged ions are called *cations,* and negatively charged ones are called *anions.*

MASS WASTING: The transfer of earth material down slopes by processes that include creep, slump, slide, flow, and fall. Also known as *mass movement.*

MINERAL: A naturally occurring, typically inorganic substance with a specific chemical composition and a crystalline structure.

ORGANIC: At one time, chemists used the term *organic* only in reference to living things. Now the word is applied to most compounds containing carbon, with the exception of carbonates (which are minerals) and oxides, such as carbon dioxide.

PRECIPITATION: In the context of chemistry, precipitation refers to the formation of a solid from a liquid.

REGOLITH: A general term describing a layer of weathered material that rests atop bedrock.

ROCK: An aggregate of minerals or organic matter, which may be consolidated or unconsolidated.

KEY TERMS CONTINUED

SEDIMENT: Material deposited at or near Earth's surface from a number of sources, most notably preexisting rock. There are three types of sediment: rocks, or clastic sediment; mineral deposits, or chemical sediment; and organic sediment, composed primarily of organic material.

SEDIMENTARY ROCK: One of the three major types of rock, along with igneous and metamorphic rock. Sedimentary rock usually is formed by the deposition, compaction, and cementation of rock that has experienced weathering. It also may be formed as a result of chemical precipitation.

SEDIMENTATION: The process of erosion, transport, and deposition undergone by sediment.

SEDIMENT LOAD: A term for the particles transported by a flowing medium of erosion (wind, water, or ice). The types of sediment load are dissolved load, suspended load, and bed load.

SEDIMENTOLOGY: The study and interpretation of sediments, including sedimentary processes and formations.

SUSPENDED LOAD: Sediment that is suspended, or floating, in the erosive medium (wind, water, or ice). Suspended load is one of three types of sediment load, along with dissolved load and bed load.

TILL: A general term for the sediments left by glaciers that lack any intervening layer of melted ice.

UNCONSOLIDATED ROCK: Rock that appears in the form of loose particles, such as sand.

WEATHERING: The breakdown of rocks and minerals at or near the surface of Earth due to physical, chemical, or biological processes.

is even more rare), it is among the most notable of placer deposits.

Of course, the fact that gold is valuable has done little to hurt, and a great deal to help, human fascination with placer gold deposits. Placer gold played a major role from the beginning of the famous California Gold Rush (1848–49), which commenced with discovery of a placer deposit by prospector James Marshall on January 24, 1848, along the American River near the town of Coloma. This discovery not only triggered a vast gold rush, as prospectors came from all over the United States in search of gold, but it also proved a major factor in the settlement of the West. Most of the miners who went to the West failed to make a fortune, of course, but instead they found something much better than gold: a gorgeous, fertile land like few places in the United States—California, a place that today holds every bit as much allure for many Americans as it did in 1848.

Despite the attention it naturally attracts, gold is far from the only placer mineral. Other placer minerals, all with a high specific gravity (density in comparison to that of water), include platinum, magnetite, chromite, native copper, zircon, and various gemstones. Nor are placer minerals found only in streams and other flowing bodies of water; wave action and shore currents can leave behind what are called beach placers. Among the notable beach placers in the world are gold deposits near Nome, Alaska, as well as zircon in Brazil and Australia, and marine gravel near Namaqualand, South Africa, which contains diamond particles.

An entirely different process can result in the formation of evaporites, minerals that include carbonates, gypsum, halites, and magnesium and potassium salts. (These specific mineral types are discussed in Minerals.) Formed when the evaporation of water leaves behind ionic, or electrically charged, chemical compounds, evaporites sometimes undergo physical processes similar to those of clastic sediment. They may even have graded bedding, meaning that the heavier materials fall to the bottom. In addition to their usefulness in industry and commerce (e.g., the use of gypsum in sheetrock for building), physical and chemical aspects of evaporites also provide scientists with considerable information regarding the past climate of an area.

WHERE TO LEARN MORE

Cherrington, Mark. *Degradation of the Land.* New York: Chelsea House, 1991.

"Erosion—Fast or Slow or None?" (Web site). <http://www.crcwater.org/issues9/19990828erosion.html>.

Middleton, Nick. *Atlas of Environmental Issues.* Illus. Steve Weston and John Downes. New York: Facts on File, 1989.

Schneiderman, Jill S. *The Earth Around Us: Maintaining a Livable Planet.* New York: W. H. Freeman, 2000.

Snedden, Robert. *Rocks and Soil.* Illus. Chris Fairclough. Austin, TX: Raintree Steck–Vaughn, 1999.

Soil Erosion and Sedimentation in the Great Lakes Region (Web site). <http://www.great-lakes.net/envt/pollution/erosion.html>.

Sediment/Soil Quality Related Links (Web site). <http://response.restoration.noaa.gov/cpr/sediment/sedlinks.html>.

State of the Land—Sedimentation and Water Quality (Web site). <http://www.nhq.nrcs.usda.gov/land/env/wq4.html>.

USDA–ARS–National Sedimentation Laboratory. United States Department of Agriculture, Agriculture Research Laboratory National Sedimentation Laboratory (Web site). <http://www.sedlab.olemiss.edu/>.

USGS Sediment Database. United States Geological Survey (Web site). <http://webserver.cr.usgs.gov/sediment/>.

Wyler, Rose. *Science Fun with Mud and Dirt.* Illus. Pat Ronson Stewart. New York: Julian Messner, 1986.

SOIL

CONCEPT

If there is anything on Earth that seems simple and ordinary, it is the soil beneath our feet. Other than farmers, people hardly think of it except when tending to their lawns, and even when we do turn our attention to the soil, we tend to view it as little more than a place where grass grows and earthworms crawl. Yet the soil is a complex mixture of minerals and organic material, built up over billions of years, and without it, life on this planet would be impossible. It is home to a vast array of species that continually process it, enriching it as they do. Nor are all soils the same; in fact, there are a great variety of soil environments and a great deal of difference between the soil at the surface and that which lies further down, closer to the bedrock.

HOW IT WORKS

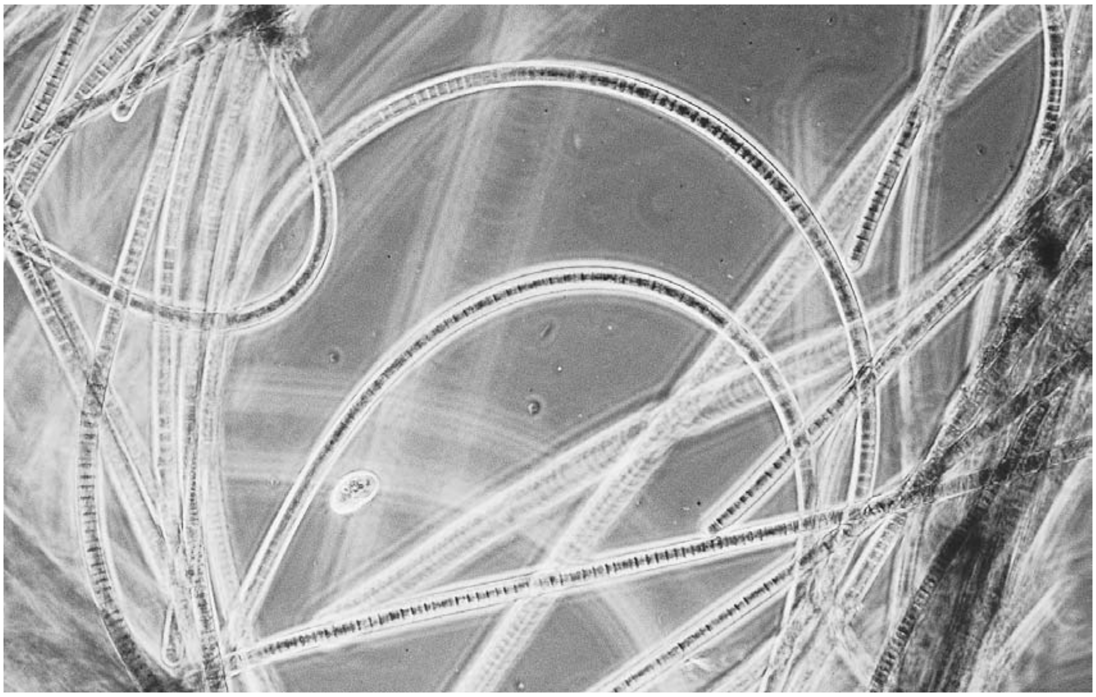
The Beginnings of Soil Formation

It has taken billions of years to yield the soil as we know it now. Over the course of these mind-boggling stretches of time, the chemical elements on Earth came into existence, and the uniformly rocky surface of the planet gradually gave way to deposits of softer material. This softer matter, the earliest ancestor of soil, became enriched by the presence of minerals from the rocks and, over a longer period, by decaying organic matter.

After its formation from a cloud of hot gas some 4.5 billion years ago, Earth was pelted by meteorites. These meteorites brought with them solid matter along with water, forming the basis for the oceans. There was no atmosphere as such, but by about four billion years ago, volcanic activity had ejected enough carbon dioxide and other substances into the air to form the beginnings of one. The oceans began to cool, making possible the earliest forms of life—that is, molecules of carbon-based matter that were capable of replicating themselves. (For more on these subjects, see Sun, Moon, and Earth and Geologic Time. On the relationship between carbon and life-forms, see Carbon Cycle.)

All of these conditions—Earth itself, an atmosphere, waters, and life-forms—went into the creation of soil. Soil has its origins in the rocks that now lie below Earth's surface, from which the rain washed minerals. For rain to exist, of course, it was necessary to have water on the planet, along with some form of atmosphere into which it could evaporate. Once these conditions had been established (as they were, over hundreds of millions of years) and the rains came down to cool the formerly molten rock of Earth's surface, a process of leaching began.

Leaching is the removal of soil particles that have become dissolved in water, but at that time, of course, there was no soil. There were only rocks and minerals, but these features of the geosphere, along with the chemical elements in the atmosphere and hydrosphere, were enough to set in motion the development of soil. While the atmosphere and hydrosphere supplied the falling rain, with its vital activity of leaching minerals from the rocks, the minerals themselves supplied additional chemical elements necessary to the formation of soil. (The chemical elements are discussed in several places, most notably Biogeochemical Cycles. See also Minerals and Rocks.)



Light micrograph of blue-green algae, an example of the simplest plant organisms that were the forerunners of life on the earth. Plant life was made possible by the leaching of potassium, calcium, and magnesium from rock, and, in turn, plant death lent organic matter to the ground to help form the basis for soil. *(© S. Stammers/Photo Researchers. Reproduced by permission.)*

THE FIRST PLANTS. Among the elements leached from the rock by the falling rains were potassium, calcium, and magnesium, all of which are essential for the growth of plant life. Thus, the foundation was laid for the first botanical forms, a fact that had several important consequences. First and most obviously, it helped set in motion the formation of the complex biosphere we have around us today. Not only did the simplest algae-like plants serve as forerunners for more complex varieties of plant and animal life to follow, but they also played a major role in the beginnings of an atmosphere breathable by animal life. As the plants absorbed carbon dioxide from their surroundings, there gradually evolved a process whereby the plant received carbon dioxide and, as a result of a chemical reaction, released oxygen.

In addition, plant life meant plant death, and as each plant died, it added just a bit more organic material—and with it nutrients and energy—to the ground. Notice the word ground as opposed to soil, which took a long, long time to form from the original rock and mineral material. Indeed, the processes we are describing here did not take shape over the course of centuries or millennia but over whole eons—the longest phases of geologic time, stretching for half a billion years or more (see Geologic Time). Only around the beginning of the present eon, the Phanerozoic, more than 500 million years ago, did soil as such begin to take shape.

What Is Soil?

As the soil began to form, processes of weathering, erosion, and sedimentation (see the entries Erosion and Sediment and Sedimentation) slowly added to the soil buildup. Today the soil forms a sheath over much of the solid earth; just inches deep or nonexistent in some places, it is many feet deep in others. It separates the planet's surface from its rocky interior and brings together a number of materials that contribute to and preserve life.

Though its origins lie in pulverized rock and decayed organic material, soil looks and feels like neither. Whether brown, red, or black, moist or dry, sandy or claylike, it is usually fairly uniform within a given area, a fact for which the organisms living in it can be thanked. Under the surface of the soil live bacteria, fungi, worms, insects, and other creatures that continually churn through it and process its chemical contents.

A filter for water and a reservoir for air, soil provides a sort of stage on which the drama of an ecosystem (a community of mutually interdependent organisms) is played out. It receives rain and other forms of precipitation, which it filters through its layers, replenishing the groundwater supplies. This natural filtration system, sometimes augmented by a little human ingenuity, is amazingly efficient for leaching out harmful microorganisms and toxins at relatively low levels. (Thus, for instance, septic tank drainage systems process wastewater, with the help of soil, before returning it to the water table.)

By collecting rainwater, soil also gives the rain a place to go and thus helps prevent flooding. Water is not the only substance it stores; soil also collects air, which accounts for a large percentage of its volume. Thus, oxygen is made available to the roots of plants and to the large populations of organisms living underground. The creatures that live in the soil also die there, providing organic material that decays along with a vast collection of dead organisms from aboveground: trees and other plants as well as dead animals—including humans, whose decomposed bodies eventually become part of the soil as well.

FACTORS THAT INFLUENCE SOIL

The processes that formed soil over the eons and that continue to contribute to the soil under our feet today are similar to those by which sedimentary rock is formed. Sedimentary rocks, such as shale and sandstone, have their origins in the deposition, compaction, and cementation of rock that has experienced weathering. Added to this is organic material derived from its ecosystem—for example, fossilized remains of animals.

Both sedimentary rock and soil are made up of sediment, which originates from the weathering, or breakdown, of rock. Weathered remains of rocks ultimately are transported by forces of erosion to what is known as a depositional environment, a location where they are sedimented. (See Sediment and Sedimentation for more about these processes.) The nature of the "parent material," or the rock from which the soil is derived, ranks among five key factors influencing the characteristics of soil in a given environment. The others are climate, living organisms, topography, and time.

PARENT MATERIAL, CLIMATE, AND ORGANISMS. Minerals, such as feldspars and micas, react strongly to natural acids carried by rain and other forms of water; therefore, when these minerals are present in the rock that makes up the parent material, they break apart quite easily into small fragments. On the other hand, a mineral that is harder—for example, quartz—will break into larger pieces of clastic, or rock, sediment. Thus, the parent material itself has a great deal to do with the initial grain of the sediment that will become soil, and this in turn influences such factors as the rate at which water leaches through it.

The release of chemical compounds and elements from minerals in weathering provides plants with the nutrients they need to grow, setting in motion the first of several steps whereby living organisms take root in, and ultimately contribute to, the soil. As the plant dies, it leaves behind material to feed decomposers, such as bacteria and fungi. The latter organisms play a highly significant role in the biogeochemical cycles whereby certain life-sustaining elements are circulated through the various earth systems.

In addition, still-living plants provide food to animals, which, when they die, likewise will become one with the soil. This is achieved through the process of decomposition, aided not only by decomposers but by detritivores as well. The latter, of which earthworms are a great example, are much more complex organisms than the typically single-cell decomposers. Detritivores consume the remains of plant and animal life, which usually contains enzymes and proteins far too complex to benefit the soil in their original state. By feeding on organic remains, detritivores cycle these complex chemicals through their systems, causing them to undergo chemical reactions that result in the breakdown of their components. As a result, simple and usable nutrients are made available to the soil.

TOPOGRAPHY AND TIME. Then there is the matter of topography, or what one might call landscape—the configuration of Earth's surface, including its relief or elevation. Soil at the top of a hill, for instance, is liable to experience considerable leaching and loss of nutrients. On the other hand, if soil is located in a basin area, it is likely to benefit from the vitamins and minerals lost to soils at higher elevations, which lose these nutrients through leaching and erosion.

In addition, topography influences the presence or absence of organic material, which is vital if the soil is to sustain plant life. Organic matter in mountainous areas accounts for only 1% to 6% of the soil composition, while in wet lowland regions it may constitute as much as 90% of soil content. Because erosion tends to bring soil, water, and organic material from the highlands to the lowlands, it is no wonder that lowlands are almost always more fertile than the mountains that surround them.

Finally, time is a factor in determining the quality of soil. As with everything else that either is living or contains living things, soil goes through a progression from immaturity to a peak to old age. In the earth sciences, age often is measured not in years, which is an absolute dating method, but by the relative dating technique of judging layers, beds, or strata of earth materials. (For more about studying rock strata as well as relative dating techniques, see Stratigraphy.)

Leaves on the forest floor are an example of humus, a component of the A horizon, or topsoil. *(© Michael Hubrich/Photo Researchers. Reproduced by permission.)*

REAL-LIFE APPLICATIONS

Layers in the Soil

If you dig down into the dirt of your backyard, you will see a miniature record of your regions's geologic history over the past few million years. Actually, most homes in urban areas and suburbs today have yards made of what is called fill dirt—loose earth that has been moved into place by a backhoe or some other earthmoving mechanism. Even though the mixed quality of fill dirt makes it difficult to discern the individual strata, the soil itself tells a tale of the long ages of time that it took to shape it.

Better than a modern fill-dirt yard, of course, would be a sample taken from an older community. Here, too, however, human activities have intervened: people have dug in their yards and holes have been filled back up, for instance, thus altering the layers of soil from what they would have been in a natural state. To find a sample of soil layers that exists in a fully natural state, it might be necessary to dig in a woodland environment.

In any case, anyone with a shovel and a piece of ground that is reasonably untouched—that is, that has not been plowed up recently—can become an amateur soil scientist. Soil scientists study soil horizons, or layers of soil that lie parallel to the surface of Earth and which have built up over time. These layers are distinguished from one another by color, consistency, and composition. A cross-section combining all or most of the horizons that lie between the surface and bedrock is called a soil profile. The most basic division of layers is between the A, B, and C horizons, which differ in depth, physical and chemical characteristics, and age.

TOPSOIL. At the top is the A horizon, or topsoil, in which humus—unincorporated, often partially decomposed plant residue—is mixed with mineral particles. Technically, humus actually constitutes something called the O horizon, the topmost layer. Examples of humus would be leaves piled on a forest floor, pine straw that covers a bare-dirt area in a yard, or grass residue that has fallen between the blades of grass on a lawn. In each case, the passage of time will make the plant materials one with the soil.

Owing to its high organic content, the soil of the A horizon may be black, or at least much darker than the soil below it. Between the A and B horizons is a noticeable layer called the *E horizon,* the depth of which is a function of the particulars in its environment, as discussed earlier. In rough terms, topsoil could be less than a foot (0.3 m) deep, or it could extend to a depth of 5 ft. (1.5 m) or more.

In any case, the E horizon, known also as the eluviation or leaching layer, is composed primarily of sand and silt, built up as water has leached down through the soil. The sediment of the E horizon is nutrient-poor, because its valuable mineral content has drained through it to the B horizon. (The E horizon is just one of several layers aside from the principal A, B, and C layers. We will mention only a few of these here, but soil scientists include several other horizons in their classification system.)

SUBSOIL, REGOLITH, BEDROCK. The appearance and consistency of the soil change dramatically again as we reach the B horizon. No longer is the earth black, even in the most organically rich environments; by this point it is more likely to exhibit shades of brown, since organic material has not reached this far below the surface. Yet subsoil, which is the consistency of clay, is certainly not poor in nutrients; on the contrary, it contains abundant deposits of iron, aluminum oxides, calcium carbonate, and other minerals, leached from the layers above it.

The rock on the C horizon is called regolith, a general term for a layer of weathered material that rests atop bedrock. Neither plant roots nor any other organic material penetrate this deeply, and the deeper one goes, the more rocky the soil. At a certain depth, it makes more sense to say that there is soil among the rocks rather than rocks in the soil.

Beneath the C horizon lies the R horizon, or bedrock. As noted earlier, depths can vary. Bedrock might be only 5–10 ft. deep (1.5–3 m), or it might be half a mile deep (0.8 km) or perhaps even deeper. Whatever the depth, it is here that the solid earth truly becomes solid, and for this reason builders of skyscrapers usually dig down to the bedrock to establish foundations there.

LIFE BENEATH THE SURFACE

The ground beneath our feet—that is, the topmost layer, the A horizon—is full of living things. In fact, there are more creatures below Earth's surface than there are above it. The term creatures in this context includes microorganisms, of which there might be several billion in a sample as small as an acorn. These include decomposers, such as bacteria and fungi, which feed on organic matter, turning fresh leaves and other material into humus. In addition, both bacteria and algae convert nitrogen into forms usable by plants in the surrounding environment (see Nitrogen Cycle).

WORMS. We cannot see bacteria, of course, but almost anyone who has ever dug in the dirt has discovered another type of organism: worms. These slimy creatures might at first seem disgusting, but without them our world could not exist as it does. As they burrow through soils, earthworms mix organic and mineral material, which they make available to plants around them. They also may draw leaves deep into their middens, or burrows, thus furnishing the soil with nutrients from the surface. In addition, earthworms provide the extraordinarily valuable service of aerating the soil, or supplying it with air: by churning up the soil continuously, they expose it to oxygen from the surface and allow air to make its way down below as well.

Nor are these visible, relatively large worms the only ones at work in the soil. Colorless worms called nematodes, which are only slightly larger than microorganisms, also live in the soil, performing the vital function of processing organic material by feeding on dead plants. Some, however, are parasites that live off the roots of such crops as corn or cotton.

ANTS AND LARGER CREATURES. Likewise there are "bad" and "good" ants. The former build giant, teeming mounds and hills that rise up like sores on the surface of the ground, and some species have the capacity to sting, causing welts on human victims. But a great number of ant species perform a positive function for the environment: like earthworms, they aerate soil and help bring oxygen and organic material from the surface while circulating soils from below.

In some areas, much larger creatures call the soil home. Among these creatures are moles, who live off earthworms and other morsels to be found beneath the surface, including grubs (insect larvae) and the roots of plants. As with ants and earthworms, by burrowing under the ground, they help loosen the soil, making it more porous and thus receptive both to moisture and air. Other large burrowing creatures include mice, ground squirrels, and prairie dogs. They typically live in dry areas, where they perform the valuable function of aerating sandy, gravelly soil.

Soils and Environments

In discussing our imaginary journey through the depths of the soil, it has been necessary to use vague terms concerning depths: "less than a foot," for instance. The reason is that no solid figures can be given for the depth of the soil in any particular area, unless those figures are obtained by a soil scientist who has studied and measured the soil.

Depth is just one of the ways that the soil may vary from one place to another. Earlier we mentioned five factors that affect the character of the soil: parent material, climate, living organisms, topography, and time. These factors determine all sorts of things about the soil—most of all, its ability to support varied life-forms. Collectively, these five factors constitute the environment in which a soil sample exists.

POOR SOILS. A desert environment might be one of immature soil, defined as a sample that has only A and C horizons, with no B horizon between them. On the other hand, the soil in rainforests suffers from just the opposite condition: it has gone beyond maturity and reached old age, when plant growth and water percolation have removed most of its nutrients.

Whether in the desert or in the rainforest, soils near the equator tend to be the "oldest," and this helps explain why few equatorial regions are noted for their agricultural productivity, even though they enjoy otherwise favorable weather for growing crops. Soils there have been leached of nutrients and contain high levels of iron oxides that give them a reddish color. Moreover, red soil is never good for growing crops: the ancient Egyptians referred to the deserts beyond their realm as "the red land," while their own fertile Nile valley was "the black land."

RAINFORESTS. If soil is so poor at the equator, why do equatorial regions such as the Congo or the Amazon River valley in Brazil support the dense, lush rain-forest ecosystems for which they are noted? The answer is that the abundance of organic material at the surface of the soil continually replenishes its nutrient content. The rapid rate of decay common in warm, moist regions further supports the process of renewing minerals in the ground.

The burrowing prairie dog helps aerate sandy, gravelly soil in dry areas. *(© Rich Kirchner/Photo Researchers. Reproduced by permission.)*

This also explains why the clearing of tropical rainforests, an issue that environmentalists called to the world's attention in the 1990s, is a serious problem. When the heavy jungle canopy of tall trees is removed, the heat of the sun and the pounding intensity of monsoon rains fall directly on ground that the canopy would normally protect. With the clearing of trees and other vegetation, the animal life that these plants support also disappears, thus removing organisms whose waste products and bodies would have decayed eventually and enriched the soil. Pounded by heat and water and without vegetation to resupply it, the soil in an exposed rainforest becomes hard and dry.

DESERTS. In deserts the soil typically comes from sandstone or shale parent material, and the lack of abundant rainfall, vegetation, or

KEY TERMS

A HORIZON: Topsoil, the uppermost of the three major soil horizons.

AERATE: To make air available to soil.

ATMOSPHERE: In general, an atmosphere is a blanket of gases surrounding a planet. Unless otherwise identified, however, the term refers to the atmosphere of Earth, which consists of nitrogen (78%), oxygen (21%), argon (0.93%), and other substances that include water vapor, carbon dioxide, ozone, and noble gases such as neon (0.07%).

B HORIZON: Subsoil, beneath topsoil and above regolith.

BEDROCK: The solid rock that lies below the C horizon, the deepest layer of soil.

BIOGEOCHEMICAL CYCLES: The changes that particular elements undergo as they pass back and forth through the various earth systems and particularly between living and nonliving matter. The elements involved in biogeochemical cycles are hydrogen, oxygen, carbon, nitrogen, phosphorus, and sulfur.

BIOSPHERE: A combination of all living things on Earth—plants, mammals, birds, reptiles, amphibians, aquatic life, insects, viruses, single-cell organisms, and so on—as well as all formerly living things that have not yet decomposed.

C HORIZON: Regolith, which lies between subsoil and bedrock and constitutes the bottommost of the soil horizons.

DECOMPOSERS: Organisms that obtain their energy from the chemical breakdown of dead organisms as well as from animal and plant waste products. The principal forms of decomposer are bacteria and fungi.

DECOMPOSITION REACTION: A chemical reaction in which a compound is broken down into simpler compounds or into its constituent elements. In the earth system, this often is achieved through the help of detritivores and decomposers.

DETRITIVORES: Organisms that feed on waste matter, breaking organic material down into inorganic substances that then can become available to the biosphere in the form of nutrients for plants. Their function is similar to that of decomposers, but unlike decomposers—which tend to be bacteria or fungi—detritivores are relatively complex organisms, such as earthworms or maggots.

ECOSYSTEM: A community of interdependent organisms along with the inorganic components of their environment.

EROSION: The movement of soil and rock due to forces produced by water, wind, glaciers, gravity, and other influences. In most cases, a fluid medium, such as air or water, is involved.

FILL DIRT: Loose earth that has been moved into place by a backhoe or some other earthmoving machine, usually as part of a large construction project.

animal life gives the soil little in the way of organic sustenance. For this reason, the A horizon level is very thin and composed of light-colored earth. Then, of course, there are desert areas made up of sand dunes, where conditions are much worse, but even the best that deserts have to offer is not very good for sustaining abundant plant life.

KEY TERMS CONTINUED

GEOSPHERE: The upper part of Earth's continental crust, or that portion of the solid earth on which human beings live and which provides them with most of their food and natural resources.

HUMUS: Unincorporated, often partially decomposed plant residue that lies at the top of soil and eventually will decay fully to become part of it.

HYDROSPHERE: The entirety of Earth's water, excluding water vapor in the atmosphere but including all oceans, lakes, streams, groundwater, snow, and ice.

LEACHING: The removal of soil materials that are in solution, or dissolved in water.

ORGANIC: At one time chemists used the term *organic* only in reference to living things. Now the word is applied to most compounds containing carbon, with the exception of carbonates (which are minerals), and oxides, such as carbon dioxide.

PARENT MATERIAL: Mineral fragments removed from rocks by means of weathering. Along with organic deposits, these form the basis for soil.

REGOLITH: A general term describing a layer of weathered material that rests atop bedrock.

SEDIMENT: Material deposited at or near Earth's surface from a number of sources, most notably preexisting rock. There are three types of sediment: rocks, or clastic sediment; mineral deposits, or chemical sediment; and organic sediment, composed primarily of organic material.

SEDIMENTARY ROCK: One of the three major types of rock, along with igneous and metamorphic rock. Sedimentary rock typically has its basis in the deposition, compaction, and cementation of rock that has experienced weathering, though it also may be formed as a result of chemical precipitation. Organic sediment also may be a part of sedimentary rock.

SEDIMENTATION: The process of erosion, transport, and deposition undergone by sediment.

SOIL HORIZONS: Layers of soil, parallel to the surface of Earth, that have built up over time. They are distinguished from one another by color, consistency, and composition.

SOIL PROFILE: A cross-section combining all or most of the soil horizons that lie between Earth's surface and the bedrock below it.

TOPOGRAPHY: The configuration of Earth's surface, including its relief as well as the position of physical features.

WEATHERING: The breakdown of rocks and minerals at or near the surface of Earth due to physical, chemical, or biological processes.

Only those species that can endure a limited water supply—for example, the varieties of cactus that grow in the American Southwest—are able to survive. But lack of water is not the only problem. Desert subsoils often contain heavy deposits of salts, and when rain or irrigation adds water to the topsoil, these salts rise. Thus, watering desert topsoil can make it a worse environment for growth.

RICH SOILS. In striking contrast to the barren soil of the deserts and the potentially barren soil of the rainforest is the rich earth that lies beneath some of the world's most fertile crop-producing regions. On the plains of the midwestern United States, Canada, and Russia, the soil is black—always a good sign for growth. Below this rich topsoil is a thick subsoil that helps hold in moisture and nutrients.

The richest variety of soil on Earth is alluvial soil, a youngish sediment of sand, silt, and clay transported by rivers. Large flowing bodies of water, such as the Nile or Mississippi, pull soil along with them as they flow, and with it they bring nutrients from the regions through which they have passed. These nutrients are deposited by the river in the alluvial soil at its delta, the place where it enters a larger body of water—the Mediterranean Sea and the Gulf of Mexico, respectively. Hence the delta regions of both rivers are extremely fertile.

WHERE TO LEARN MORE

Bial, Raymond. *A Handful of Dirt.* New York: Walker, 2000.

Bocknek, Jonathan. *The Science of Soil.* Milwaukee, WI: Gareth Stevens, 1999.

Canadian Soil Information System (Web site). <http://sis.agr.gc.ca/cansis/>.

Gardner, Robert. *Science Projects About the Environment and Ecology.* Springfield, NJ: Enslow Publishers, 1999.

Scheiderman, Jill S. *The Earth Around Us: Maintaining a Livable Planet.* New York: W. H. Freeman, 2000.

Snedden, Robert. *Rocks and Soil.* Illus. Chris Fairclough. Austin, TX: Raintree Steck–Vaughn, 1999.

Soil Association (Web site). <http://www.soilassociation.org>.

Soil Science Society of America (Web site). <http://www.soils.org/>.

USDA-NRCS National Soil Survey Center (Web site). <http://www.statlab.iastate.edu/soils/nssc/>.

World Soil Resources (Web site). <http://www.nhq.nrcs.usda.gov/WSR/Welcome.html>.

SOIL CONSERVATION

CONCEPT

With the rise of the environmentalist movement in the 1960s and afterward, it has become common to speak of conserving natural resources such as trees or fossil fuels. Yet long before humans recognized the need to make responsible use of things taken from the ground, they learned to conserve the ground itself—that is, the soil. This was a hard-won lesson: failure to conserve soil has turned many a fertile farmland into temporary dust bowl or even permanent desert. Techniques such as crop rotation aid in conservation efforts, but communities continue to face hazards associated with the soil. There is, for instance, the matter of leaching, the movement of dissolved substances through the soil, which, on the one hand, can benefit it but, on the other hand, can rob it of valuable nutrients. Issues of soil contamination also raise concerns that affect not just farmers but the population as a whole.

HOW IT WORKS

Billions of Years in the Making

Earth's present wealth of soil is the result of hundreds of millions of years' worth of weathering, erosion, and sedimentation. Once, long ago, there was no soil, only rock, and it took eons' worth of weathering to dislodge particles of those rocks. These rocks, when combined with organic materials, became the basis for soil, but before the soil could even begin to take shape, a number of things had to fall into place. Chief among these was the formation of something that, at first glance, at least, does not seem to have a great deal of bearing on the soil: the atmosphere.

In combination with water in the hydrosphere (e.g., streams and rivers) as well as water in the form of evaporated moisture and precipitation in the air itself, the blanket of gases we call our atmosphere has been essential to the formation and sustenance of Earth's soil. This importance goes beyond the obvious point that rain transports water to the soil, thus making possible the abundance of plant life that grows in it. Rain, of course, is of unquestionable importance, but it is only one of several factors associated with the atmosphere (including the water vapor it contains) that have a role in shaping soil as we know it.

To move weathered rocks from highlands to lowlands, where they can become sediment and eventually begin to form soil, it is necessary to subject the rocks themselves to a process of erosion. And erosion—aside from erosion caused by gravity, which usually is considered weathering—can take place only when an atmosphere exists, along with water in the air and on the land. The chief agents of erosion are wind, water (both flowing and in the form of precipitation), and frozen water in the form of icy glaciers, all of which depend on an atmosphere or water or both (see Glaciology).

Erosion transports not only rock sediment but organic material as well. Together, these two ingredients are as essential to making soil as tea bags and water are to making tea. Obviously, the greater the organic content, the richer the soil, and here again the air plays a part. It is important that deeper layers of soil receive a supply of air from the surface to sustain the life of subterranean organisms, who not only process nutrients through the soil but (by their burrowing activities) also aerate it, or make air available to it.

A Product of Its Environment

Soil, like most people, is a product of the environment in which it was formed. That environment has five major influencing factors: the nature of the "parent material," or the rock from which the soil was derived; the local climate; the presence of living organisms; local topography; and the passage of time.

Specific classes of mineral break apart in characteristic ways, and the size of the pieces into which the original weathered rock is broken has a great deal to do with the character of the soil that it forms. This does not mean, however, that relatively large rock pieces necessarily will yield the worst soils, since erosive forces will continue to work on the rock, pulling out its nutrient-rich mineral wealth and gradually acting to break it apart.

As for climate, it is clear that rain and sun are essential for the growth of plant matter, but, of course, too much of either or both is harmful. (See Soil for a discussion of soils in rainforests.) Plants aid the soil by dying and feeding it with more organic material, but they are not the only types of organism in the soil. Indeed, the soil constitutes an ecosystem in and of itself, a realm rich in biodiversity, in which various biogeochemical cycles are played out, and through which energy flows as part of the operation of the larger Earth system.

The underground world teems with creatures ranging from bacteria to moles and prairie dogs (in some regions), each of which fulfills a function. These functions include aerating the soil by burrowing; processing material though ingestion and elimination of waste, thus converting compounds into nutrients that the soil can use; and mixing organic material with minerals. Organisms' final contribution to the soil comes when they die, as their bodies become material that feeds the earth through decomposition.

Topography, or elevation, plays a major role in making possible erosion, itself a process that can be either beneficial or detrimental. The question of whether it is one or the other may be a matter of perspective, or rather elevation. From the standpoint of lowland areas, which receive the wealth of the upland areas in the form of nutrient-rich runoff carried by gravity or flowing media, such as wind or water, erosion is a good thing. Matters do not look as good from the viewpoint of the mountains, which lose much of their best soil to low-lying areas.

The influence of time in shaping soils—as well as much else about the soil itself—can be appreciated by studying soil horizons, the various strata, or layers, of soil that lie beneath the surface. The most basic division of layers is between the A, B, and C horizons, which differ in depth and physical and chemical characteristics as well as age.

SOIL HORIZONS. Above the A horizon, or topsoil, lies humus, decomposing organic material that eventually will become soil. The A horizon itself contains a large amount of organic matter, and thus it may be black, or at least much darker than the soil below it. Between the A and B horizons is a sandy, silty later called the E horizon. Then comes the B horizon, or subsoil, which starts at a depth as shallow as 1 ft. (0.3 m) or deeper than 5 ft. (1.5 m).

Lacking a great deal of organic material but still rich in nutrients, the B horizon has a sizable impact on the A horizon. Minerals—both healthful and harmful—may rise up from the B to the A horizon, and the ability of the B horizon to hold in moisture from above greatly affects the moisture of the A horizon soil. (See Soil for a discussion of how salt deposits in the B horizon affect topsoil in deserts.) Together, A and B horizons constitute what is called the *solum,* or true soil.

The C horizon is called *regolith.* It is the home for the rocks of the parent material, which has given up much of its nutrient riches in fortifying the soil that lies above it. This far below the surface, there is no sign of plant or animal life, and below the C horizon is the R horizon, or bedrock—the top of the layers of rock and metal that descend all the way to the planet's core. Once again depths vary, with bedrock as shallow as 5–10 ft. (1.5–3 m) or as deep as 0.5 mi. (0.8 km) or more.

Differences Between Soils

The depth of the soil is a measure of wealth—wealth, that is, in terms of natural resources. A sheath over much of the solid earth, soil separates the planet's surface from its rocky interior and preserves the lives of the plants and animals

A CLOUD OF TOPSOIL IS PICKED UP BY THE WIND NEAR BOISE CITY, OKLAHOMA, DURING THE DUST BOWL OF THE 1930S. IN SOME CASES, WIND REMOVED 3-4 IN. (7.6-10.6 CM) OF TOPSOIL, TURNING ACREAGE THAT ONCE RIPPLED WITH WHEAT INTO A DESERTLIKE WASTELAND. *(AP/Wide World Photos. Reproduced by permission.)*

that live on and in it. It receives rain and other forms of precipitation, which it filters through its layers, as we discuss later, in the context of leaching. Thus, it not only provides water to organisms above and below its surface but also helps prevent flooding by acting as a reservoir.

A great deal of soil's volume is air, for which it also acts as a reservoir. Underground creatures depend on this air and also help circulate it by burrowing. This circulation, in turn, provides oxygen to the roots of plants and makes the soil more hospitable to growth. Even though soil performs these and other life-preserving functions, it would be a mistake to assume that all soils are the same. In fact, the U.S. Department of Agriculture has identified 11 major soil orders, each of which is divided into suborders, groups, subgroups, families, and series.

The specificity of soil types, as reflected in the identification and naming of soil series, illustrates the complexity of what at first seems a very simple thing. In fact, soils can be extremely specific, with names that reflect local landmarks. If soils share enough similarities, they are grouped together in a soil series, but it is safe to say that there are thousands of individual soil types on Earth.

CONSERVING SOIL

On a broad level, there are certain types of environment more or less favorable to the formation of rich soil. Some of these types are discussed in the essay Soil, and specific examples of environmental problems are provided later in this essay. Yet almost any environment can become unfavorable to plant growth if proper soil-conservation procedures are not observed.

The phrase soil conservation refers to the application of principles for maintaining the productivity and health of agricultural land by control of wind- and water-induced soil erosion. For the remainder of this essay, we examine the dangers involved in such erosion and the use of measures to prevent it. In so doing, we give the matter of soil conservation a somewhat larger scope than the preceding definition might suggest. Since soil affects the world far beyond farms, it seems only fitting to approach it not as a concern merely of agriculture but of the environment in general.

Erosion is spoken of here in a general sense, but for a more in-depth discussion of erosive processes, see Erosion. Mass Wasting examines dramatic erosion-related phenomena, such as landslides. Biogeochemical Cycles contains some discussion of erosion, inasmuch as it helps circulate life-sustaining chemical elements throughout the various earth systems. Indeed, it is important to remember that erosion is not always negative in its results; on the contrary, it is a valuable process by which landforms are shaped. The erosive processes we explore here, however, generally contribute to the loss of soil health and productivity.

REAL-LIFE APPLICATIONS

The Dust Bowl

When people mismanage agricultural lands or when natural forces otherwise conspire to destroy soil, the results can be devastating. One of the most dramatic examples occurred in what came to be known as the dust bowl. This was the name given to a wide area covering Texas, Oklahoma, Kansas, and even agricultural parts of Colorado during the years 1934 and 1935. Over the course of a few months, once-productive farmlands turned into worthless fields of stubble and dust, good for almost nothing and highly vulnerable to violent wind erosion.

And wind erosion came, scattering vast quantities of soil from the Great Plains of the Midwest to the Atlantic Seaboard. The classic 1939 film *The Wizard of Oz* sets its fantastic, otherworldly story against this backdrop, and to viewers in the late 1930s the tornado that swept Dorothy from her Kansas farmland into the world of Oz was all too real. The only difference was that no magical adventure awaited victims of the real-life tornadoes and other windstorms.

The fate of the dust bowl farmers, many of whom lost everything, was dramatized in the novel *The Grapes of Wrath* by John Steinbeck in 1939 as well as in the acclaimed motion picture that followed a year later. A perhaps equally eloquent tribute appeared in the form of the American photographer Dorothea Lange's photographs of dust bowl refugees. The images etched by Lange are unforgettable: in one a woman stares into the distance, her face a landscape of despair, as her children huddle next to her, their eyes hidden from the camera. In another a man, obviously exhausted from months or years of overwork, hardship, and fear, sits behind the wheel of a truck, gazing somewhere beyond the camera lens. Like the woman, he seems to be looking into a future that offers scant hope.

CAUSES OF THE DUST BOWL. What happened? The sad fact is that in the years leading up to the early 1930s, the future dust bowl farmlands had seemed remarkably productive, and farmers continued to be pleasantly surprised, year after year, at the abundant yields they could draw from each field. In fact, farmers were unwittingly preparing the way for vast erosion by overcultivating the land and not taking proper steps to preserve its moisture against drought. This was particularly unfortunate because farmers in the 1930s had long known about the principle of crop rotation as a means of giving the soil a rest in order to restore its nutrients. Yet the farmers of the plains tried to push their crops to yield more and more, and for a time it worked, though at great future expense to the land.

One is tempted to see in the agricultural world of the U.S. Midwest parallels to the foolhardy attitude that, just a few years earlier, created a boom on Wall Street, followed by the devastating stock market crash of October 29, 1929, that ushered in the Great Depression. Certainly the ravages of the dust bowl, when they came, were particularly unwelcome in a land already reeling from several years of widespread unemployment and a sagging economy. And though there was no cause-effect relationship between the Wall Street crash and the dust bowl, there is no question that both were brought about in large part by a lack of planning for the future and by a naive belief that it is possible to get "something for nothing"—that is, to get more out of the world (whether the world of finances or the natural world) than one puts into it.

In some places farmers alternated between wheat cultivation and livestock grazing on particular plots of land. Thus, the hooves of the cattle damaged the soil, which had been weakened by raising wheat. The land was therefore ready to become the site of a full-fledged natural disaster, and, at the height of the depression, natural disaster came in the form of high winds. The winds in some cases removed topsoil as much as 3–4 in. (7–10 cm) thick. Dunes of dust as tall as 15–20 ft. (4.6–6.1 m) formed, turning acreage that once had rippled with wheat into desertlike wastelands.

Erosion Control in Action

Today the farmlands of the plains states long since have recovered, and American farmers have benefited from the lessons learned in the dust bowl. Out of the dust bowl years came the establishment, in 1935, of the Soil Conservation Service, a federal agency charged with implementing erosion-control practices. (The Soil Conservation Service was the predecessor of the modern-day Natural Resources Conservation Service.) In the wake of the legislation creating the agency, signed into law by President Franklin D. Roosevelt (1882–1945), states passed laws creating nearly 3,000 local soil conservation districts.

If one passes through agricultural lands today, one is likely to see signs identifying the local conservation district. Even more important, the lands themselves are a testament to principles put into practice as an outgrowth of the dust bowl years. For instance, instead of alternating one year of wheat with one year in which a field lies fallow, or unused, farmers in the dust bowl region discovered that a three-year cycle of wheat, sorghum, and fallow land worked much better. They also planted trees to serve as barriers against wind.

EROSION CONTROL LEGISLATION. Concerns over soil conservation in America did not end with the dust bowl. As United States farm production soared in the 1970s, American farms enjoyed such a great surplus that U.S. farmers increasingly began to sell their crops overseas—most notably, to the Soviet Union. While some Americans were upset to see the farmers of the Midwest selling wheat to the Communists in Moscow, others saw in this act a testament to the failure of the Soviet agricultural system and to the strength of U.S. farming. In the wake of these increased exports, farmers were encouraged to cultivate even marginal croplands to increase profits, thus heightening the vulnerability of their lands to erosion.

What followed was not another dust bowl, however; instead, the experience of the 1970s and 1980s shows just how much American farmers, legislators, and others had learned from the 1930s. Environmental activists in the 1970s, concerned over water quality, helped return public interest to the problem of soil erosion. They called attention to the flow of nutrients from croplands into water resources, most notably leaching of nitrogen and phosphorus that choked lakes with eutrophication (see Biogeochemical Cycles). As a result of public concerns over these and related issues, Congress in 1977 passed the Soil and Water Resources Conservation Act, mandating the conservation of soil, water, and other resources on private farmlands and other properties.

In 1985 the Food Security Act further served to encourage steps toward the reduction of soil erosion. Some 45 million acres (18 million hectares) of land vulnerable to erosion were removed from intensive cultivation by the act. The legislation also forbade the conversion of rangelands into agricultural fields, which would have raised great potential for erosion and depletion of already vulnerable soil. In addition, the act required farmers to develop and maintain practices for the control of erosion on lands susceptible to that threat.

BARRIER AND COVER. Soil-conservation practices fall under two headings: barrier and cover. Under the barrier approach, various structures act as a wall against water runoff, wind, and the movement of soil. Among such structures are banks, hedgerows, walls of earth or other materials, and silt fences such as one sees at construction sites. The cover approach is devoted to the idea of maintaining a heavy soil cover of living and dead plant material. This is achieved through the use of mulch, cover crops, and other techniques.

Local governments and property owners in nonagricultural lands often apply both the cover and barrier approaches, planting trees as well as grass not simply to beautify the land but also to hold the soil in place. Land has to have some sort of vegetative protection to stand between it and the forces of wind and water erosion, and the two approaches together serve to protect soil against nature's onslaught.

Leaching

Like erosion, leaching—the movement of dissolved substances with water percolating through soil—can be both positive and negative. For any plot of land, assuming the rate of water input is greater than the rate of water loss through evaporation, water has to go somewhere, so it leaves the site by moving downward. Eventually it either reaches the deep groundwater or passes through subterranean springs to flow into the surface waters of streams, rivers, and lakes.

Along the way, the leached water carries all sorts of dissolved substances, ranging from nutrients to contaminants. The threat of the latter has led to widespread concern in the United States over the leaching of toxins into water supplies, and in 1980 this concern spurred a massive piece of legislation called CERLA (Comprehensive Environmental Response, Compensation, and Liability Act), better known as Superfund. Six years later, in 1986, Congress updated CERLA with the Superfund Amendments and Reauthorization Act. These laws provided for a vast array of measures directed toward environmental cleanup, including the removal of chemicals and other toxins in soil.

Drastic measures such as those outlined in CERLA and other legislation may be required for the cleanup of artificial materials introduced into soils and groundwater. But for human waste and other more natural forms of toxin, nature itself is able to achieve a certain amount of cleanup on its own. In a septic-tank system, used by people who are not connected to a municipal sewage system, bacteria process wastes, removing a great deal of their toxic content in the tank itself. The wastewater leaves the tank and passes through a filtration system, in which the water leaches through layers of gravel and other filters that help remove more of its harmful content. As the wastewater percolates from the filtration system through the soil (usually well below the A horizon by this point), it is purified further before it enters the groundwater supply.

Not only does leaching help purify the water that passes through the soil, it also is an important part of the soil-formation process, inasmuch as it passes nutrients to the depths of the A horizon and into the B horizon. Its ability to pass along nutrients is not always beneficial, and in some ecosystems, large amounts of dissolved nitrogen are lost to soil as a result of leaching. In such a situation, soil typically is fertilized with nitrate, a form of the element with which soil often has difficulty binding (see Nitrogen Cycle). For this reason, nitrate tends to leach easily, leading to an overabundance of nitrogen in the lower levels of the soil and in the groundwater. This condition, known as nitrogen saturation, can influence the eutrophication of waters (see Biogeochemical Cycles for an explanation of eutrophication) and can cause the decline and death of trees on the surface.

Desertification

Much of North Africa lies under the cover of a vast desert, the Sahara. By far the world's largest desert, the Sahara today spreads across some 3.5 million sq. mi. (9.06 million sq km), an area larger than the continental United States. Only about 780 acres (316 hectares) of it, or little more than 1 sq. mi. (2.6 sq km), is fertile. The rest is mostly stone and dry earth with scattered shrubs—and, here and there, the rolling sand dunes typically used to depict the Sahara in movies.

Given the forbidding moonscape of the Sahara today, it might be surprising to learn that just 8,000 years ago—the blink of an eye in terms of geologic time—it was a region of flowing rivers and lush valleys. For thousands of years it served as a home to many cultures, some of them quite advanced, to judge from their artwork. Though they left behind an extraordinary record in the form of their rock-art paintings and carvings, which show an understanding of realistic representation that would not be matched until the time of the Greeks, the identity of the early Saharan peoples themselves remains largely a mystery.

Instead of identifying them by the name of a nationality or empire, archaeologists divide the phases of the early Saharan culture according to a set of four names that collectively tell the story of the region's progressive transformation into a desert. First was the Hunter period, from about 6000 to about 4000 B.C., when a Paleolithic, or Old Stone Age, people survived by hunting the many wild animals then available in the region. Next came the Herder period, from about 4000 to 1500 B.C. As their name suggests, these people maintained herds of animals and also practiced basic agriculture.

As the Sahara became drier and drier, however, there were no more herds. Egyptians began bringing in domesticated horses to cross the desert: hence the name of the Horse period (*ca.* 1500–*ca.* 600 B.C.) By about 600 B.C., not even horses could survive in the forbidding climate. There was only one creature that could survive: the hardy, seemingly inexhaustible camel. Thus began the Camel era, which continues to the present day.

ATTEMPTS TO CONTROL DESERTIFICATION. As with the dust bowl, the first question one wants to ask when confronted

A camel caravan in the Sahara. The world's largest desert, it covers 3.5 million sq. mi. (9 million sq km), but 8,000 years ago this was a region of lush valleys and flowing rivers. *(© Tom Hollyman/Photo Researchers. Reproduced by permission.)*

with a story such as that of the Sahara, is "What happened?" The answer is much more complex, just as the effects of desertification—the slow transformation of ordinary lands to desert—are much more permanent than those of the erosion associated with the dust bowl. Desertification does not always result in what people normally think of as a desert. It is rather a process that contributes toward making a region more dry and arid, and because it is usually gradual, it can be reversed in some cases. But doing so represents a vast challenge.

In 1977 the United Nations (UN), in the form of the UN Conference on Desertification in Nairobi, Kenya, set out to address the spread of the Sahara into the Sahel, an arid region that stretches south of the desert. Some 700 delegates from almost 100 countries adopted a number of measures designed to halt the spread of desertification in that region and others by the year 2000.

Even though there have been some successes, the Sahel region today remains a blighted area where famine is common, and this state of affairs is not entirely the result of the natural causes addressed in the conference's resolutions. Poor government management and a near-constant state of civil war in such countries as Ethiopia have played at least as important a role in spreading famine as nature itself. During the 1980s, in fact, the government of Ethiopia (at that time a Marxist-Leninist state) deliberately withheld food supplies, shipped to it from the West, as a way of exerting pressure on rebel factions and other groups it wished to subdue.

THE EXAMPLE OF IRAQ. The arid regions of Iraq provide another example of how human influences can result in desertification. Once that country, known in ancient times as Mesopotamia, was among the greenest and most lush places in the known world. For this reason, historians today use the name Fertile Crescent to describe an arc from the deltas of the Tigris and Euphrates rivers in Mesopotamia to the mouth of the Nile in Egypt. Today, of course, Iraq is mostly a dust-colored land of bare trees and brush.

What happened? Agricultural mismanagement certainly played a role, as did the simple exhaustion of the soil by some 6,000 years of human civilization. Indeed, since the Fertile Crescent was perhaps the first area settled by agricultural societies long before the beginning of full-fledged civilization as such in about 3500

KEY TERMS

A HORIZON: Topsoil, the uppermost of the three major soil horizons.

AERATE: To make air available to soil.

B HORIZON: Subsoil, beneath topsoil and above regolith.

BEDROCK: The solid rock that lies below the C horizon, the deepest layer of soil.

BIOGEOCHEMICAL CYCLES: The changes that particular elements undergo as they pass back and forth through the various earth systems and particularly between living and nonliving matter. The elements involved in biogeochemical cycles are hydrogen, oxygen, carbon, nitrogen, phosphorus, and sulfur.

C HORIZON: Regolith, which lies between subsoil and bedrock and constitutes the bottommost of the soil horizons.

DECOMPOSERS: Organisms that obtain their energy from the chemical breakdown of dead organisms as well as from animal and plant waste products. The principal forms of decomposer are bacteria and fungi.

DECOMPOSITION REACTION: A chemical reaction in which a compound is broken down into simpler compounds or into its constituent elements. In the earth system, this often is achieved through the help of detritivores and decomposers.

DETRITIVORES: Organisms that feed on waste matter, breaking organic material down into inorganic substances that then can become available to the biosphere in the form of nutrients for plants. Their function is similar to that of decomposers; however, unlike decomposers—which tend to be bacteria or fungi—detritivores are relatively complex organisms, such as earthworms or maggots.

ECOSYSTEM: A community of interdependent organisms along with the inorganic components of their environment.

EROSION: The movement of soil and rock due to forces produced by water, wind, glaciers, gravity, and other influences. In most cases, a fluid medium, such as air or water, is involved.

EUTROPHICATION: A state of heightened biological productivity in a body of water, which is typically detrimental to the ecosystem in which it takes place. Eutrophication can be caused by an excess of nitrogen or phosphorus in the form of nitrates and phosphates, respectively.

HUMUS: Unincorporated, often partially decomposed plant residue that lies at the top of soil and eventually will decay fully to become part of it.

B.C., it is safe to say that the region has been under cultivation for several thousand years longer—perhaps 8,000 or even 10,000 years. Direct human action and malice also may have played a role: some historians believe that the Mongols, during their brutal invasion in the 1250s, so badly devastated the farmlands and irrigation channels of Iraq that the land never recovered.

SOME CAUSES OF DESERTIFICATION. With regard to human involvement in the desertification process, it is not necessary for a society to be advanced agriculturally to do long-term damage to the soil. The Pueblan

KEY TERMS CONTINUED

LANDFORM: A notable topographical feature, such as a mountain, plateau, or valley.

LEACHING: The removal of soil materials that are in solution, or dissolved in water.

MASS WASTING: The transfer of earth material down slopes by processes that include creep, slump, slide, flow, and fall. Also known as *mass movement.*

MINERAL: A naturally occurring, typically inorganic substance with a specific chemical composition and a crystalline structure.

ORGANIC: At one time chemists used the term *organic* only in reference to living things. Now the word is applied to most compounds containing carbon, with the exception of carbonates (which are minerals) and oxides, such as carbon dioxide.

PARENT MATERIAL: Mineral fragments removed from rocks by means of weathering. Along with organic deposits, these fragments form the basis for soil.

REGOLITH: A general term describing a layer of weathered material that rests atop bedrock.

SEDIMENT: Material deposited at or near Earth's surface from a number of sources, most notably preexisting rock. There are three types of sediment: rocks, or clastic sediment; mineral deposits, or chemical sediment; and organic sediment, composed primarily of organic material.

SEDIMENTATION: The process of erosion, transport, and deposition undergone by sediment.

SOIL CONSERVATION: The application of principles for maintaining the productivity and health of agricultural land by control of wind- and water-induced soil erosion. The term also may be applied more broadly to encompass the maintenance and protection of nonagricultural soils.

SOIL HORIZONS: Layers of soil, parallel to the surface of the earth, which have built up over time. These layers are distinguished from one another by color, consistency, and composition.

SOIL PROFILE: A cross-section combining all or most of the soil horizons that lie between Earth's surface and the bedrock below it.

WEATHERING: The breakdown of rocks and minerals at or near the surface of Earth due to physical, chemical, or biological processes.

culture of what is now the southwestern United States depleted an already dry and vulnerable region after about A.D. 800 by removing its meager stands of mesquite trees. And though human causes, in the form of either mismanagement or deliberate damage, have contributed toward desertification, sometimes nature itself is the driving force.

Long-term changes in rainfall or general climate as well as water erosion and wind erosion such as caused the dust bowl can turn a region into a permanent desert. An ecosystem may survive short-term drought, but if soil is forced to go too long without proper moisture, it sets in motion a chain reaction in which plant life dwindles and, with it, animal life as well. Thus, the soil is denied the fresh organic material necessary to its continued sustenance, and a slow, steady process of decline begins.

WHERE TO LEARN MORE

Bear, Firman E., H. Wayne Pritchard, and Wallace E. Akin. *Earth: The Stuff of Life*. Norman: University of Oklahoma Press, 1986.

Bocknek, Jonathan. *The Science of Soil.* Milwaukee, WI: Gareth Stevens, 1999.

Bright Edges of the World: The Earth's Evolving Drylands (Web site). <http://www.nasm.edu/ceps/drylands/drylands.html>.

Cherrington, Mark. *Degradation of the Land.* New York: Chelsea House, 1991.

"Desertification." United States Geological Survey (Web site). <http://pubs.usgs.gov/gip/deserts/desertification/>.

Gardner, Robert. Science *Projects About the Environment and Ecology.* Springfield, NJ: Enslow Publishers, 1999.

Natural Resources Conservation Service (Web site). <http://www.nrcs.usda.gov/>.

Pittman, Nancy P. *From the Land.* Washington, DC: Island Press, 1988.

Soil and Water Conservation Society (Web site). <http://www.swcs.org/>.

"Voices from the Dust Bowl: The Charles L. Todd and Robert Sonkin Migrant Worker Collection, 1940–1941." Library of Congress (Web site). <http://memory.loc.gov/ammem/afctshtml/tshome.html>.

GEOCHEMISTRY

BIOGEOCHEMICAL CYCLES

CONCEPT

Of the 92 elements produced in nature, only six are critical to the life of organisms: hydrogen, carbon, nitrogen, oxygen, phosphorus, and sulfur. Though these elements account for 95% of the mass of all living things, their importance extends far beyond the biosphere. Hydrogen and oxygen, chemically bonded in the form of water, are the focal point of the hydrosphere, while oxygen and nitrogen form the bulk of the atmosphere. All six are part of complex biogeochemical cycles in which they pass through the biosphere, atmosphere, hydrosphere, and geosphere. These cycles circulate nutrients through the soil into plants, microbes, and animals, which return the elements to the earth system through chemical processes that range from respiration to decomposition.

HOW IT WORKS

The Elements

An element is a substance composed of a single type of atom, meaning that it cannot be broken down chemically to make a simpler substance. They are listed on the periodic table of elements, a chart that renders them in order of their atomic numbers, or the number of protons in the nucleus of the atom. The elements we want to discuss in the context of biogeochemical cycles are all low in atomic number, starting with hydrogen, which has just one proton in its nucleus. In the following list, the elements are cited by atomic number along with their chemical symbols, or the abbreviation by which they are known to chemists.

Elements Involved in Biogeochemical Cycles

- 1. Hydrogen (B)
- 6. Carbon (C)
- 7. Nitrogen (N)
- 8. Oxygen (O)
- 15. Phosphorus (P)
- 16. Sulfur (S)

Given the fact, as noted earlier, that 92 elements appear in nature, it should come as no surprise that the highest atomic number for any naturally occurring element is 92, for uranium. Beyond uranium there are about two dozen artificially created elements, but they are of little interest outside the realm of certain specialties in chemistry and physics. The naturally occurring elements are the ones that matter to the earth sciences, and of these elements, only a handful play a significant role.

In the essays Minerals and Economic Geology, other elements—most notably silicon—are discussed with regard to their importance in forming minerals, rocks, and ores. Though they are critical to Earth's systems, elements other than the six discussed here play no role in biogeochemical cycles. Indeed, it is a fact of the physical sciences that not all elements are created equal: certainly, the universe is not divided evenly 92 ways, with equal amounts of all elements. In fact, hydrogen and helium account for 99% of the mass of the entire universe.

ABUNDANCE. On Earth the ratios are quite different, however. Oxygen and silicon constitute the preponderance of the known mass of Earth's crust, while nitrogen and oxygen form the overwhelming majority of the atmosphere. Hydrogen is proportionally much, much less abundant on Earth than in the universe as a whole, but owing to its role in forming water, a substance essential to the sustenance of life, it is unquestionably of great significance.

The two following lists provide rankings for the abundance of the six elements discussed in this essay. The first table shows their ranking and share in the entire known mass of the planet, including the crust, living matter, the oceans, and atmosphere. The second shows their relative abundance and ranking in the human body.

Abundance of Selected Elements on Earth (Ranking and Percentage)

- 1. Oxygen (49.2%)
- 9. Hydrogen (0.87%)
- 12. Phosphorus (0.11%)
- 14. Carbon (0.08%)
- 15. Sulfur (0.06%)
- 16. Nitrogen (0.03%)

Abundance of Selected Elements in the Human Body (Ranking and Percentage)

- 1. Oxygen (65%)
- 2. Carbon (18%)
- 3. Hydrogen (10%)
- 4. Nitrogen (3%)
- 6. Phosphorus (1%)
- 9. Sulfur (0.26%)

Several things are interesting about these figures. First and most obviously, there is the fact that the ranking of all these elements (with the exception of oxygen) is relatively low in the total known elemental mass of Earth, whereas their ranking is much, much higher within the human body. This is significant, given the fact that these elements are all essential to the lives of organisms.

Furthermore, note that it does not take a great percentage to constitute an "abundant" element: even nitrogen, with its 0.03% share of Earth's total known mass, still is considered abundant. The presence of the vast majority of elements on Earth is measured in parts per million (ppm) or even parts per billion (ppb).

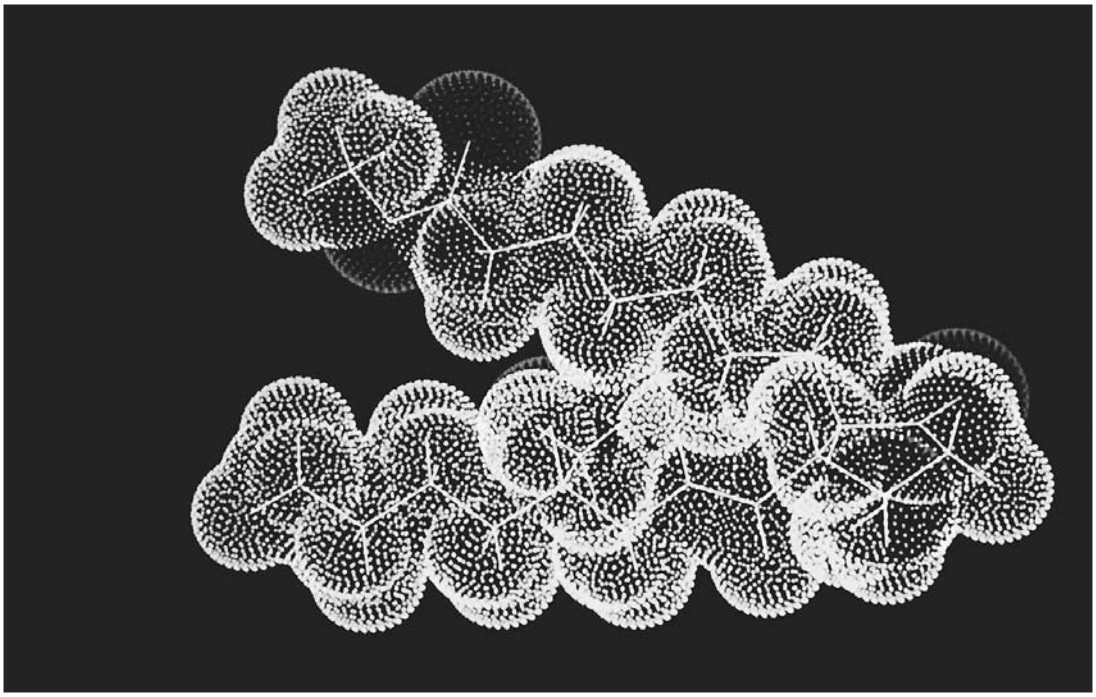
Chemistry and Geochemistry

In a general sense, chemistry can be defined as an area of the physical sciences concerned with the composition, structure, properties, and changes of substances, including elements, compounds, and mixtures. This definition unites the phases in the history of the development of the discipline, from early modern times—when it arose from alchemy, a set of mystical beliefs based on the idea that ordinary matter can be perfected—to modern times. Our modern understanding of chemistry, however, is quite different from the model of chemistry that prevailed until about 1800, a difference that relates to a key discovery: the atom.

This change, in fact, should be described not in terms of a discovery so much as the development of a model. Long before chemists and physicists comprehended the structure of the atom, they developed an understanding of all chemical substances as composed of atomic units, each representing one and only one element. Until the development of this model—thanks to a number of chemists, most notably, John Dalton (1766–1844) of England, Antoine Lavoisier (1743–1794) of France, and Amedeo Avogadro (1776–1856) of Italy—chemistry was concerned primarily with mixing potions and observing their effects. Thanks to the atomic model, chemists never again would confuse mixtures with chemical compounds.

CHEMICAL REACTIONS. The difference between a mixture and a compound goes to the heart of the distinction between physics and chemistry. A mixture, such as coffee, is the result of a physical process—in this case, the heating of water and coffee beans—and the result does not have a uniform chemical structure. On the other hand, a compound results from chemical reactions between atoms, which form enormously powerful bonds in the process of joining to create a molecule. A molecule is the basic particle of a compound, just as an atom is to an element. It should be noted that some elements, such as nitrogen, typically appear in diatomic form, that is, two atoms bond to form a molecule of nitrogen.

A substance may undergo physical changes without experiencing any alteration in its underlying structure; on the other hand, a chemical reaction makes a fundamental change to the substance. In a chemical reaction, a substance may experience a change of state (i.e., from solid to liquid or gas) without undergoing any physical process of being heated or cooled by an outside source. Chemical reactions involve the breaking of bonds between atoms in a molecule and the formation of new bonds. As a result, an entirely new



Molecular structure. A compound is formed by chemical reactions between atoms, which join together to make molecules. *(© Blair Seitz/Photo Researchers. Reproduced by permission.)*

substance is created—something that could never be achieved through mere physical processes.

REAL-LIFE APPLICATIONS

Geochemistry

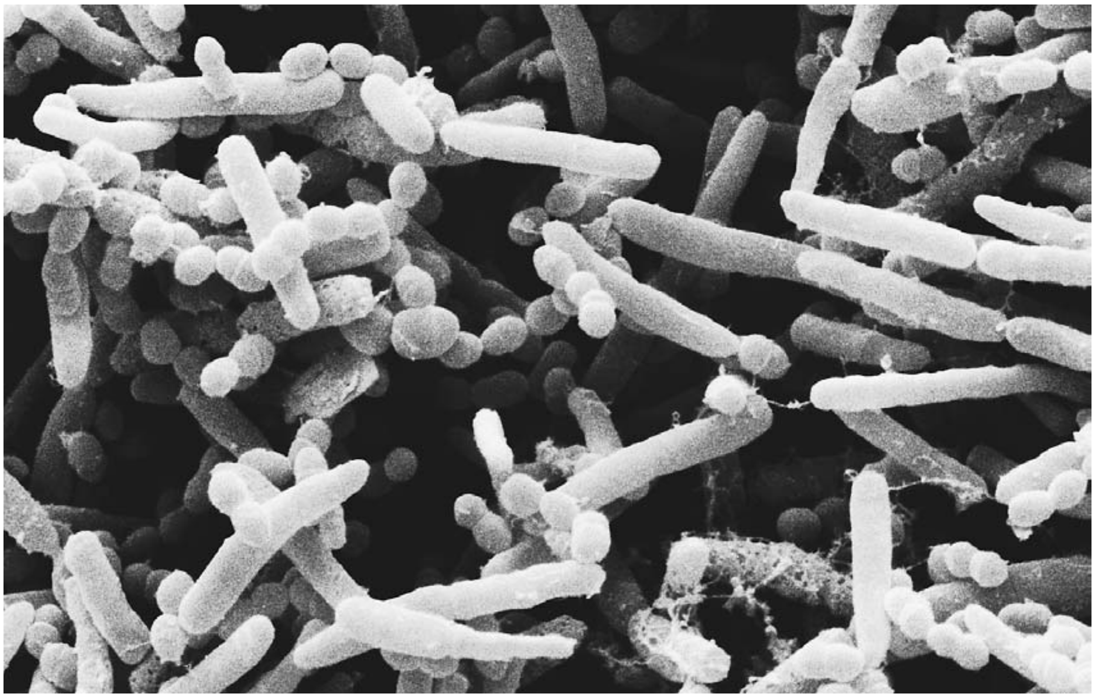
Just as geochemistry is a branch of the geologic sciences that weds physics and geology, so there is a geologic subdiscipline, geochemistry, in which chemistry and the geologic sciences come together. Geochemistry is concerned with the chemical properties and processes of Earth—in particular, the abundance and interaction of chemical elements and their isotopes. (Isotopes are atoms that have an equal number of protons, and hence are of the same element, but differ in their number of neutrons. Isotopes may be either stable or unstable, in which case they are subject to the emission of high-energy particles. Some elements have numerous stable isotopes, others have only one or two, and some have none.)

Before the mid–twentieth century, geochemistry had a relatively limited scope, confined primarily to the identification of elements in rocks and minerals and the determination of the relative abundance of those elements. Since that time, however, this subdiscipline has come to encompass many other concerns, particularly those discussed in the present context. Geochemistry today focuses on such issues as the recycling of elements between the various sectors of the earth system, especially between living and nonliving things.

Biogeochemical Cycles and Earth Systems

The changes that a particular element undergoes as it passes back and forth through the various earth systems, and particularly between living and nonliving matter, are known as biogeochemical cycles. The four earth systems involved in these cycles are the atmosphere, the biosphere (the sum of all living things as well as formerly living things that have not yet decomposed), the hydrosphere (the entirety of Earth's water except for vapor in the atmosphere), and the geosphere. The last of these spheres is defined as the upper part of Earth's continental crust, or that portion of the solid earth on which human beings live and which provides them with most of their food and natural resources.

Carbon, for instance, is present in all living things on Earth. Hence, the phrase *carbon-based life-form,* a cliché found in many an old science-



BACTERIA BEGIN THE PROCESS OF DECOMPOSITION OF ORGANIC WASTE, BREAKING DOWN PLANT MATTER AND CONVERTING IT INTO COMPOST. *(© Scimat/Photo Researchers. Reproduced by permission.)*

fiction movie, is actually a redundancy: *all* lifeforms contain carbon. In the context of the physical sciences, *organic* refers to all substances that contain carbon, with the exception of oxides, such as carbon dioxide and carbon monoxide, and carbonates, which are found in minerals. Still, carbon circulates between the organic world and the inorganic world, as when an animal exhales carbon dioxide.

The carbon cycle is of such importance to the functioning of Earth that it is discussed separately (see Carbon Cycle). So, too, is the nitrogen cycle (see Nitrogen Cycle), whereby nitrogen passes between the soil, air, and biosphere as well as the hydrosphere. The hydrosphere, as noted earlier, is based on a single substance, water, created by the chemical bonding of hydrogen and oxygen, and it is likewise discussed in detail elsewhere (see Hydrologic Cycle).

Despite the emphasis here on carbon in the biosphere, nitrogen in the geosphere, and hydrogen and oxygen in the hydrosphere, it should be noted that biogeochemical cycles involving these four elements take them through all four "spheres." The same is true of sulfur, whose biogeochemical cycle is discussed later in this essay. On the other hand, phosphorus, also discussed later, is present in only three of Earth's systems; it plays little role in the atmosphere.

DECOMPOSERS AND DETRITIVORES

Most biogeochemical cycles involve a special type of chemical reaction known as decomposition, and for this to take place, agents of decomposition—known as decomposers and detritivores—are essential. Decomposition occurs when a compound is broken down into simpler compounds or into its constituent elements. This is achieved primarily by decomposers, organisms that obtain their energy from the chemical breakdown of dead organisms as well as from animal and plant waste products.

The principal forms of decomposer are bacteria and fungi. These creatures carry enzymes, which they secrete into the materials they consume, breaking them down chemically before taking in the products of this chemical breakdown. They thus take organic matter and render it in inorganic form, such that later it can be taken in again by plants and returned to the biosphere.

Detritivores are much more complex organisms, but their role is similar to that of decomposers. They, too, feed on waste matter, breaking this organic material down into inorganic substances that then can become available to the biosphere in the form of nutrients for plants. Examples of detritivores are earthworms and maggots. As discussed in Energy and Earth, detritivores are key players in the food web, the set of nutritional interactions—sometimes called a food chain—between living organisms.

Phosphorus and the Phosphorus Cycle

There are a few elements that were known in ancient or even prehistoric times, examples being gold, iron, lead, and tin. The vast majority, on the other hand, have been discovered since the beginning of the modern era, and the first of them was phosphorus, which is also the first element whose discoverer is known.

In 1674 the German alchemist Hennig Brand (*ca.* 1630–*ca.* 1692) was searching for the philosopher's stone, a mythical substance that allegedly would turn common or base metals into gold. Convinced that he would find this substance in the human body, Brand evaporated water from a urine sample and burned the precipitate (the solid that remained) along with sand. The result was a waxy, whitish substance that glowed in the dark and reacted violently with oxygen. Brand named it phosphorus, a name derived from a Greek term meaning "light-bearer."

Owing to its high reactivity with oxygen, phosphorus is used in the production of safety matches, smoke bombs, and other incendiary devices. It is also important in various industrial applications and in fertilizers. In fact, ancient humans used phosphorus without knowing it when they fertilized their crops with animal bones.

PHOSPHATES. In the early 1800s, chemists recognized that the critical component in bones was phosphorus, which plants use in photosynthesis—the biological conversion of energy from the Sun into chemical energy (see Energy and Earth). With this discovery came the realization that phosphorus would make an even more effective fertilizer when treated with sulfuric acid, which makes it soluble, or capable of being dissolved, in water. This compound, known as superphosphate, can be produced from phosphates, a type of mineral.

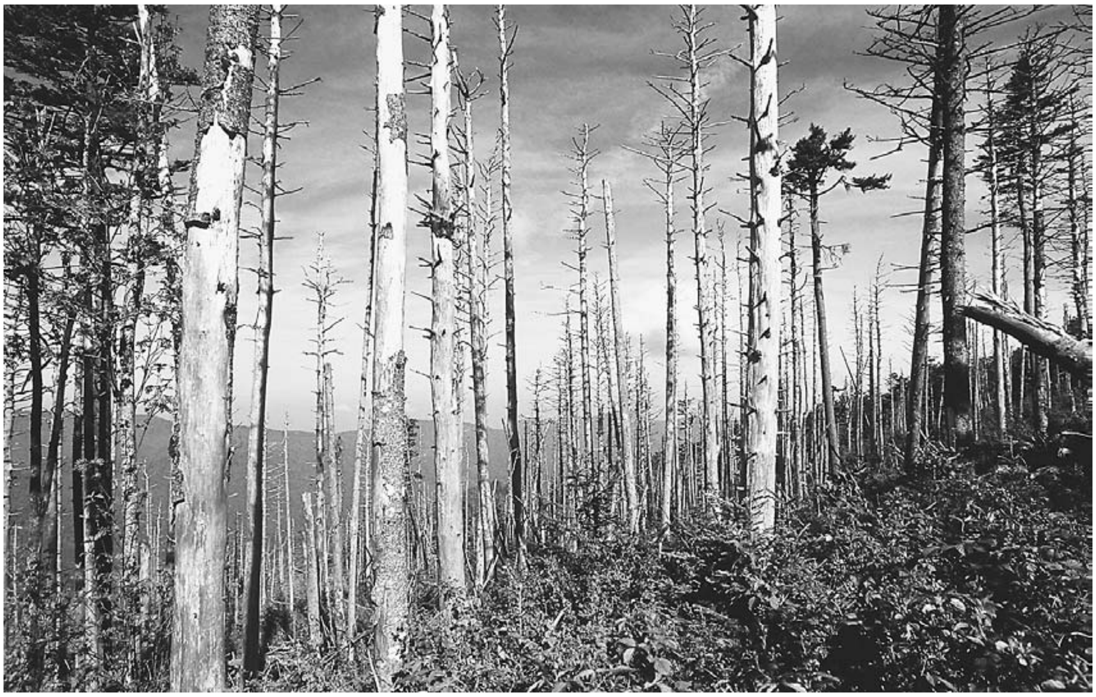
Phosphates represent one of the eight major classes of mineral (see Minerals). All phosphates contain a characteristic formation, PO_4, which is bonded to other elements or compounds—for example, with aluminum in aluminum phosphate, or $AlPO_4$. Phosphorus fertilizer is typically calcium phosphate, known as bone ash, the most important industrial mineral (see Economic Geology) produced from phosphorus. Another significant phosphate is sodium phosphate, used in dishwashing detergents. In fact, phosphates once played a much larger role in the detergent industry—with disastrous consequences, as we shall see.

THE PHOSPHORUS CYCLE. The majority of phosphorus in the earth system is located in rocks and deposits of sediment, from which it can be removed by one of three processes: weathering, the breakdown of rocks and minerals at or near the surface of Earth as the result of physical, chemical, or biological processes (see Erosion, Sediment and Sedimentation); leaching, the removal of soil materials that are in solution, or dissolved in water; and mining.

Phosphorus is highly reactive, meaning that it is likely to bond with other elements, and for this reason it often is found in compounds. Microorganisms absorb insoluble phosphorus compounds (ones that are incapable of being dissolved) and, through the action of acids within the microorganisms, turn them into soluble phosphates. Algae and other green plants absorb these phosphates and, in turn, are eaten by animals. When they die, the animals release the phosphates back into the soil.

As with all elements, the total amount of phosphorus on Earth stays constant, but the distribution of it does not. Some of the phosphorus passes from the geosphere into the biosphere, but the majority of it winds up in the ocean. It may find its way into sediments in shallow waters, in which case it continues to circulate, or it may be taken to the deep parts of the seas, in which case it is likely to be deposited for the long term.

Fish absorb particles of phosphorus, and thus some of the element returns to dry land through the catching and consumption of seafood. In addition, guano, or dung, from birds that live in an ocean environment (e.g., seagulls) also returns portions of phosphorus to the terrestrial environment. Nonetheless, geochemists



A stand of fir trees shows the devastating effects of acid rain, which is created when sulfuric acid mixes with moisture in the atmosphere. *(© Will and Demi McIntyre/Photo Researchers. Reproduced by permission.)*

believe that phosphorus is being transferred steadily to the ocean, from whence it is not likely to return. It is for this reason that phosphorus-based fertilizers are important, because they feed the soil with nutrients that otherwise would be continually lost.

EUTROPHICATION. To be sure, phosphorus, in the proper quantities, is good for the environment. But as with medicine or any other beneficial substance, if a little is good, that does not mean that a lot is necessarily better. In the case of phosphorus, an overabundance of the element in the environment can lead to a phenomenon called eutrophication, a state of heightened biological productivity in a body of water. One of the leading causes of eutrophication (from a Greek term meaning "well nourished") is a high rate of nutrient input, in the form of phosphates or nitrates, a nitrogen-oxygen compound (see Nitrogen Cycle).

As a result of soil erosion, fertilizers make their way into bodies of water, as does detergent runoff in wastewater. Excessive phosphates and nitrates stimulate growth in algae and other green plants, and when these plants die, they drift to the bottom of the lake or other body of water. There, decomposers consume the remains of the plants and, in the process, also use oxygen that otherwise would be available to fish, mollusks, and other forms of life. As a result, those species die off, to be replaced by others that are more tolerant of lowered oxygen levels—for example, worms. Needless to say, the outcome of eutrophication is devastating to the lake's ecosystem.

During the 1960s, Lake Erie—one of the Great Lakes on the U.S.-Canadian border—became an example of eutrophication gone mad. As a result of high phosphate concentrations, Erie's waters were choked with plant and algae growth. Fish were unable to live in the water, the beaches reeked with the smell of decaying algae, and Erie became widely known as a "dead" body of water. This situation led to the passage of new environmental standards and pollution controls by both the United States and Canada, whose governments acted to reduce drastically the phosphate content in fertilizers and detergents. Lake Erie proved to be an environmental success story: within a few decades the lake once again teemed with life.

Sulfur and the Sulfur Cycle

If there is any element that can be said to have a bad image—and a falsely bad one at that—it is sulfur. As everyone "knows," sulfur has a foul smell, and this smell, combined with its com-

KEY TERMS

ALCHEMY: A set of mystical beliefs based on the idea that ordinary matter can be perfected. Though it was a pseudoscience, alchemy, which flourished in the late Middle Ages, was a forerunner of scientific chemistry.

ATMOSPHERE: In general, an atmosphere is a blanket of gases surrounding a planet. Unless otherwise identified, however, the term refers to the atmosphere of Earth, which consists of nitrogen (78%), oxygen (21%), argon (0.93%), and other substances that include water vapor, carbon dioxide, ozone, and noble gases such as neon (0.07%).

ATOMIC NUMBER: The number of protons in the nucleus of an atom. Since this number is different for each element, elements are listed on the periodic table in order of atomic number.

BIOGEOCHEMICAL CYCLES: The changes that particular elements undergo as they pass back and forth through the various earth systems and particularly between living and nonliving matter. The elements involved in biogeochemical cycles are hydrogen, oxygen, carbon, nitrogen, phosphorus, and sulfur.

BIOSPHERE: A combination of all living things on Earth—plants, mammals, birds, amphibians, reptiles, aquatic life, insects, viruses, single-cell organisms, and so on—as well as all formerly living things that have not yet decomposed.

CARNIVORE: A meat-eating organism.

CHEMICAL BONDING: The joining through electromagnetic force of atoms that sometimes, but not always, represent more than one chemical element. The result is the formation of a molecule.

CHEMICAL SYMBOL: A one-letter or two-letter abbreviation for the name of an element.

COMPOUND: A substance made up of atoms of more than one element chemically bonded to one another.

DECOMPOSERS: Organisms that obtain their energy from the chemical breakdown of dead organisms as well as from animal and plant waste products. The principal forms of decomposers are bacteria and fungi.

DECOMPOSITION REACTION: A chemical reaction in which a compound is broken down into simpler compounds or into its constituent elements. In the earth system, this often is achieved through the help of detritivores and decomposers.

DETRITIVORES: Organisms that feed on waste matter, breaking organic material down into inorganic substances that then can become available to the biosphere in the form of nutrients for plants. Their function is similar to that of decomposers; however, unlike decomposers—which tend to be bacteria or fungi—detritivores are relatively complex organisms, such as earthworms or maggots.

ECOSYSTEM: A term referring to a community of interdependent organisms along with the inorganic components of their environment.

ELEMENT: A substance made up of only one kind of atom. Unlike compounds, elements cannot be broken chemically into other substances.

KEY TERMS CONTINUED

EUTROPHICATION: A state of heightened biological productivity in a body of water, which is typically detrimental to the ecosystem in which it takes place. Eutrophication can be caused by an excess of nitrogen or phosphorus in the form of nitrates and phosphates, respectively.

FOOD WEB: A term describing the interaction of plants, herbivores, carnivores, omnivores, decomposers, and detritivores, each of which consumes nutrients and passes it along to other organisms.

GEOCHEMISTRY: A branch of the earth sciences, combining aspects of geology and chemistry, that is concerned with the chemical properties and processes of Earth—in particular, the abundance and interaction of chemical elements and their isotopes.

GEOSPHERE: The upper part of Earth's continental crust, or that portion of the solid earth on which human beings live and which provides them with most of their food and natural resources.

HERBIVORE: A plant-eating organism.

HYDROSPHERE: The entirety of Earth's water, excluding water vapor in the atmosphere but including all oceans, lakes, streams, groundwater, snow, and ice.

LEACHING: The removal of soil materials that are in solution, or dissolved in water.

MINERAL: A naturally occurring, typically inorganic substance with a specific chemical composition and a crystalline structure.

MIXTURE: A substance with a variable composition, meaning that it is composed of molecules or atoms of differing types. Compare with *compound.*

MOLECULE: A group of atoms, usually but not always representing more than one element, joined in a structure. Compounds typically are made up of molecules.

bustibility, led to the biblical association of *brimstone*—the ancient name for the element—with the fires of hell. It may come as a surprise, then, to learn that sulfur has no smell of its own. Only in combination with other elements does it acquire the offensive odor that has led to its unpleasant reputation.

An example of such a compound is hydrogen sulfide, a poisonous substance present in intestinal gas. The May 2001 *National Geographic* included two stories relating to the presence of natural hydrogen sulfide deposits on opposite sides of the earth, and in both cases the presence of these toxic fumes created unusual ecosystems.

A system of caves known as Villa Luz in southern Mexico contains some 20 underground springs that carry large quantities of hydrogen sulfide. Among the strange creatures that have made Villa Luz their home are species of fish colored bright red; the pigmentation is a result of the fact that they have to produce high quantities of hemoglobin (a component in red blood cells) to survive on the scant oxygen. The waters of the cave also are populated by microorganisms that oxidize the hydrogen sulfide and turn it into sulfuric acid, which dissolves the rock walls and continually enlarges the cave.

Thousands of miles away, in the Black Sea, explorers examining evidence of a great ancient flood like the one depicted in the Bible (see Earth, Science, and Nonscience) found an unexpected ally in the form of hydrogen sulfide. Because the Black Sea lacks the temperature differences that cause water to circulate from the bottom upward, hydrogen sulfide had gathered at the bottom and stayed there, covered by dense layers of saltwater. Oxygen could not reach the

KEY TERMS CONTINUED

OMNIVORE: An organism that eats both plants and other animals.

ORGANIC: At one time chemists used the term *organic* only in reference to living things. Now the word is applied to most compounds containing carbon, with the exception of carbonates (which are minerals) and oxides, such as carbon dioxide.

PERIODIC TABLE OF ELEMENTS: A chart that shows the elements arranged in order of atomic number along with their chemical symbols and the average atomic mass for each particular element.

PHOTOSYNTHESIS: The biological conversion of light energy (that is, electromagnetic energy) from the Sun to chemical energy in plants.

REACTIVITY: A term referring to the ability of one element to bond with others. The higher the reactivity, the greater the tendency to bond.

SEDIMENT: Material deposited at or near Earth's surface from a number of sources, most notably preexisting rock. There are three types of sediment: rocks, or clastic sediment; mineral deposits, or chemical sediment; and organic sediment, composed primarily of organic material.

SOLUBLE: Capable of being dissolved.

SOLUTION: A homogeneous mixture (i.e., one that is the same throughout) in which one or more substances is dissolved in another substance—for example, sugar dissolved in water.

WEATHERING: The breakdown of rocks and minerals at or near the surface of Earth due to physical, chemical, or biological processes.

bottom of the Black Sea, and thus wood-boring worms could not live in the toxic environment. As a result, a 1,500-year-old shipwreck had been virtually undisturbed.

THE SULFUR CYCLE. Sulfur is removed from rock by weathering, at which point it reacts with oxygen in the air to form sulfate, or SO_4. This sulfate is taken in by plants and microorganisms, which convert it to organic materials and pass it on to animals in the food web. Later, when these organisms die, decomposers absorb the sulfur from their bodies and return it to the environment. As with phosphorus, however, sulfur is being lost continually to the oceans as it drains through lakes and streams (and through the atmosphere) on its way to the sea.

In the ocean ecosystem, sulfur can take one of three routes. Some of it circulates through food webs, and some drifts to the bottom to bond with iron in the form of ferrous sulfide, or FeS. Ferrous sulfide contributes to the dark color of sediments at the bottom of the ocean. On the other hand, sulfur may be returned to the atmosphere, released by spray from saltwater. In addition, sulfur can pass into the atmosphere as the result of volcanic activity or through the action of bacteria, which release it in the form of hydrogen sulfide, the foul-smelling gas discussed earlier.

As with all biogeochemical cycles, humans play a part in the sulfur cycle, and the role of modern industrial society is generally less than favorable, as is true of most such cycles. A particularly potent example is the production of acid rain. Among the impurities in coal is sulfur, and when coal is burned (as it still is, for instance, in electric power plants), it results in the production of sulfur dioxide and sulfur trioxide—SO_2

and SO_3, respectively. Sulfur trioxide reacts with water in the air to produce sulfuric acid, or H_2SO_4. This mixes with moisture in the atmosphere to create acid rain, which is hazardous to both plant and animal life.

WHERE TO LEARN MORE

Beatty, Richard. *Sulfur.* New York: Benchmark Books, 2000.

———. *Phosphorus.* New York: Benchmark Books, 2001.

Biogeochemical Cycles (Web site). <http://www.bsi.vt.edu/chagedor/biol_4684/Cycles/cycles.html>.

General Chemistry Online (Web site). <http://antoine.fsu.umd.edu/chem/senese/101/index.shtml>.

Geochemistry on the World Wide Web (Web site). <http://www.geo.cornell.edu/geology/classes/Geochemweblinks.HTML>.

Hancock, Paul L., Brian J. Skinner. *The Oxford Companion to the Earth.* New York: Oxford University Press, 2000.

Kump, Lee R., James F. Kasting, and Robert G. Crane. *The Earth System.* Upper Saddle River, NJ: Prentice Hall, 2000.

Life and Biogeochemical Cycles (Web site). <http://essp.csumb.edu/esse/climate/climatebiogeo.html>.

Skinner, Brian J., Stephen C. Porter, and Daniel B. Botkin. *The Blue Planet: An Introduction to Earth System Science.* 2d ed. New York: John Wiley and Sons, 1999.

WebElements Periodic Table of the Elements (Web site). <http://www.webelements.com/webelements/scholar/index.html>.

THE CARBON CYCLE

CONCEPT

If a person were asked to name the element most important to sustaining life, chances are he or she would say oxygen. It is true that many living things depend on oxygen to survive, but, in fact, carbon is even more fundamental to the sustenance of life. Indeed, in a very real sense, carbon *is* life, since every living thing contains carbon and the term *organic* refers to certain varieties present in all life-forms. Yet carbon, in the form of such oxides as carbon dioxide as well as carbonates like calcium carbonate, is a vital part of the inorganic realm as well. Hence, the carbon cycle, by which the element is circulated through the biosphere, geosphere, atmosphere, and hydrosphere, is among the most complex of biogeochemical cycles.

HOW IT WORKS

Geochemistry

Chemistry is concerned with the composition, structure, properties, and changes of substances, including elements, compounds, and mixtures. Central to the discipline is the atomic model, or the idea that all matter is composed of atoms, each of which represents one and only one chemical element. An element thus is defined as a substance made up of only one kind of atom, which cannot be broken chemically into other substances. A chemical reaction involves either the bonding of one atom with another or the breaking of chemical bonds between atoms.

Geochemistry brings together geology and chemistry, though as the subdiscipline has matured in the period since the 1940s, its scope has widened to take in aspects of other disciplines and subdisciplines. With its focus on such issues as the recycling of elements between the various sectors of the earth system, especially between living and nonliving things, geochemistry naturally encompasses biology, botany, and a host of earth science subdisciplines, such as hydrology.

BIOGEOCHEMICAL CYCLES. Among the most significant areas of study within the realm of geochemistry are biogeochemical cycles. These are the changes that a particular element undergoes as it passes back and forth through the various earth systems—particularly between living and nonliving matter. As we shall see, this transition between the worlds of the living and the nonliving is particularly interesting where carbon is concerned.

Along with carbon, five other elements—hydrogen, nitrogen, oxygen, phosphorus, and sulfur—are involved in biogeochemical cycles. With the exception of phosphorus, which plays little part in the atmosphere, these elements move through all four earth systems, including the atmosphere, the biosphere (the sum of all living things as well as formerly living things that have not yet decomposed), the hydrosphere (Earth's water, except for water vapor in the atmosphere), and the geosphere, or the upper part of Earth's continental crust.

Earth systems and biogeochemical cycles are discussed in greater depth within essays devoted to those topics (see Earth Systems and Biogeochemical Cycles). Likewise, the nitrogen cycle is treated separately (see Nitrogen Cycle). The role of hydrogen and oxygen, which chemically bond

THE NAME CARBON COMES FROM THE LATIN WORD FOR CHARCOAL, CARBO. COAL HAS A WIDE VARIETY OF USES, FROM MANUFACTURING STEEL TO GENERATING ELECTRICITY. *(© Andrew J. Martinez/Photo Researchers. Reproduced by permission.)*

to form water, is discussed in the context of the hydrosphere (see Hydrologic Cycle).

ELEMENTS AND COMPOUNDS

We have referred to elements and compounds, which are essential to the study of chemistry; now let us examine them briefly before going on to the subject of a specific and very important element, carbon. An element is defined not by outward characteristics, though elements do have definable features by which they are known; rather, the true meaning of the term *element* is discernible only at the atomic level.

Every atom has a nucleus, which contains protons, or subatomic particles of positive electric charge. The identity of an element is defined by the number of protons in the nucleus: for instance, if an atom has only a single proton, by definition it must be hydrogen. An atom with six protons in the nucleus, on the other hand, is always an atom of carbon. Thus, the elements are listed on the periodic table of elements by atomic number, or the number of protons in the atomic nucleus.

ELECTRONS AND CHEMICAL REACTIONS. While protons are essential to the definition of an element, they play no role in the bonding between atoms, which usually produces chemical compounds. (The reason for this is qualified by the modifier *usually,* in that sometimes two atoms of the same element may bond as well.) Chemical bonding involves only the electrons, which are negatively charged subatomic particles that spin around the nucleus. In fact, only certain of these fast-moving particles take part in bonding. These are the valence electrons, which occupy the highest energy levels in the atom.

One might say that valence electrons are at the "outside edge" of the atom, though the model of atomic structure, considered only in the briefest form here, is far more complex than that phrase implies. In any case, elements have characteristic valence electron patterns that affect their reactivity, or their ability to bond. Carbon is structured in such a way that it can form multiple bonds, and this feature plays a significant part in its importance as an element.

When an element reacts with another, they join together, generally in a molecule (we will examine some exceptions), to form a compound. Though the atoms themselves remain intact, and an element can be released from a compound, a compound quite often has properties quite unlike those of the original elements. Carbon and oxygen are essential to sustaining life, but when a single atom of one bonds with a single atom of the other, they form a toxic gas, carbon monoxide. And whereas carbon in its elemental form is a black powder and hydrogen and oxygen are colorless, odorless gases, when bonded in the proper proportions and structure, the three create sugar.

CARBON

The name carbon comes from the Latin word for charcoal, *carbo.* In fact, charcoal—wood or other plant material that has been heated without enough air present to make it burn—is just one of many well-known substances that contain carbon. Others include coal, petroleum, and other fossil fuels, all of which contain hydrocarbons, or chemical compounds built around strings of carbon and hydrogen atoms. Graphite is pure carbon, and coke, a refined version of coal, is very nearly pure. Not everything made of carbon is black, however: diamonds, too, are pure carbon in another form.

Though carbon makes up only a small portion of the known elemental mass in Earth's crust, waters, and atmosphere—just 0.08%, or 1/1,250 of the whole—it is the fourteenth most abundant element on the planet. In the human body, carbon is second only to oxygen in abundance and accounts for 18% of the body's mass. Present in the inorganic rocks of the ground and in the living creatures above it, carbon is everywhere in the earth system.

CARBON BONDING. There are two elements noted for their ability to form long strings of atoms and seemingly endless varieties of molecules: one is carbon, and the other is silicon, directly below it on the periodic table. Just as carbon forms a vast array of organic compounds, silicon, found in a huge variety of minerals, is at the center of a large number of inorganic compounds. Yet carbon is capable of forming an even greater number of bonds than silicon. (For more about silicon and the silicates, see the entries Minerals and Economic Geology.)

Carbon is distinguished further by its high value of electronegativity, the relative ability of an atom to attract valence electrons. In addition, with four valence electrons, carbon is ideally suited to finding other elements (or other carbon atoms) with which to form chemical bonds. Normally, an element does not necessarily have the ability to bond with as many other elements as it has valence electrons, but carbon—with its four valence electrons—happens to be tetravalent, or capable of bonding to four other atoms at once. Additionally, carbon can form not just a single bond but also a double bond or even a triple bond with other elements.

A DIAMOND IS AN ALLOTROPE, A CRYSTALLINE FORM, OF CARBON. ESSENTIALLY, IT IS A HUGE MOLECULE COMPOSED OF CARBON ATOMS STRUNG TOGETHER BY COVALENT BONDS. *(© V. Fleming/Photo Researchers. Reproduced by permission.*

ALLOTROPES OF CARBON. Carbon has several allotropes—different versions of the same element distinguished by their molecular structure. The first of them is graphite, a soft material that most of us regularly encounter in the form of pencil "lead." Graphite is essentially a series of one-atom-thick sheets of carbon bonded together in a hexagonal pattern, but with only very weak attractions between adjacent sheets.

Then there is that most alluring of all carbon allotropes, diamond. Neither diamonds nor graphite, strictly speaking, are formed of molecules. Their arrangement is definite, as with a molecule, but their size is not: they simply form repeating patterns that seem to stretch on forever. Whereas graphite is in the form of sheets, a diamond is basically a huge "molecule" composed of carbon atoms strung together by what are known as covalent chemical bonds.

Graphite and diamond are both crystalline—solids in which the constituent parts have a simple and definite geometric arrangement that is repeated in all directions. (All minerals are crystalline in structure. See Minerals.) A third carbon allotrope, buckminsterfullerene, discovered in 1985 and named after the American engineer and philosopher R. Buckminster Fuller (1895–1983), is also crystalline in form.

Carbon takes yet another form, distinguished from the other three allotropes in that it is amorphous in structure—lacking a definite shape—as opposed to crystalline. Though it retains some of the microscopic structures of the plant cells in the wood from which it is made, charcoal is mostly amorphous carbon. Coal and coke are particularly significant varieties of amorphous carbon. Formed by the decay of fossils, coal was the first important fossil fuel (discussed later in this essay) used to provide heat and power to human societies.

REAL-LIFE APPLICATIONS

ORGANIC CHEMISTRY

Organic chemistry is the study of carbon, its compounds (with the exception of the carbonates and oxides mentioned earlier), and their properties. At one time chemists thought that *organic* was synonymous with living, and even as recently as the early nineteenth century, they believed that organic substances contained a supernatural "life force." Then, in 1828, the German chemist Friedrich Wöhler (1800–1882) made an amazing discovery.

By heating a sample of ammonium cyanate, a material from a nonliving source, Wöhler converted it to urea, a waste product in the urine of mammals. As he later observed, "without benefit of a kidney, a bladder, or a dog," he had turned an inorganic substance into an organic one. It was almost as though he had created life. Actually, what he had discovered was the distinction between organic and inorganic material, which results from the way in which the carbon chains are arranged.

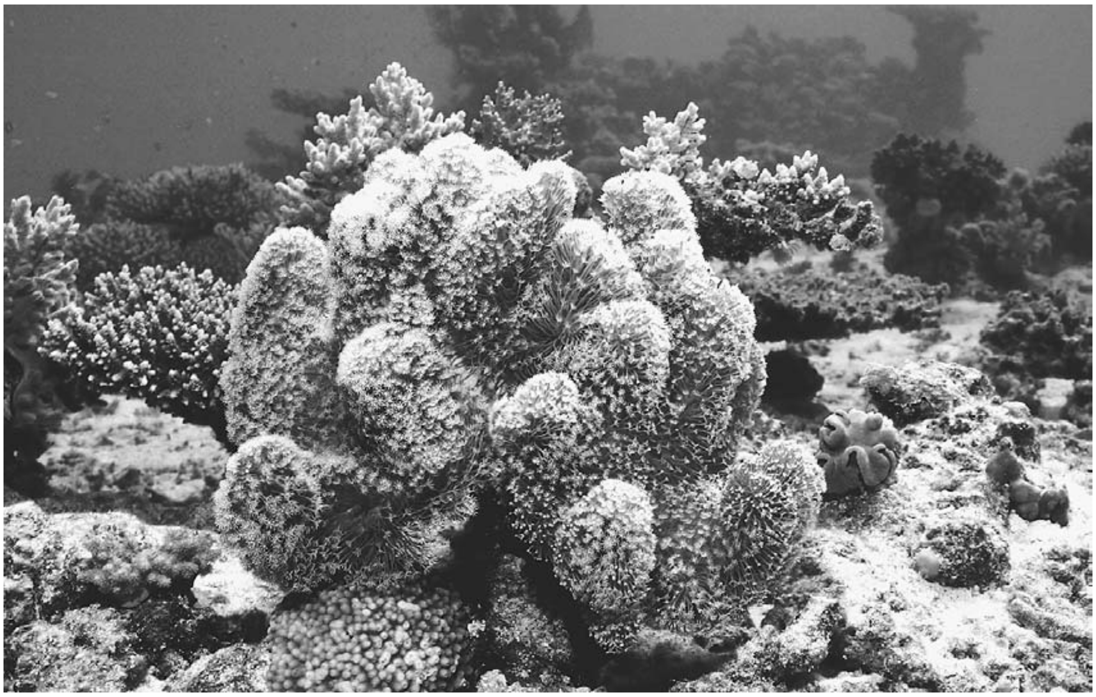
Organic chemistry encompasses the study of many things that people commonly think of as "organic"—living creatures, formerly living creatures, and the parts and products of their bodies—but it also is concerned with substances that seem quite far removed from the living world. Among these substances are rubber, vitamins, cloth, and paper, but even in these cases, it is easy to see the relationship to a formerly living organism: a rubber plant, or a tree that was cut down to make wood pulp. But it might come as a surprise to learn that plastics, which at first glance would seem completely divorced from the living world, also have an organic basis. All manner of artificial substances, such as nylon and polyester, are made from hydrocarbons.

FOSSIL FUELS. During the Mesozoic era, which began about 248.2 million years ago, dinosaurs ruled the earth; then, about 65 million years ago, a violent event brought an end to their world. The cause of this mass extinction is unknown, though it is likely that a meteorite hit the planet, sending so much dust into the atmosphere that it dramatically changed local climates, bringing about the destruction of the dinosaurs—along with a huge array of other animal and plant forms. (See Paleontology for more on this subject.)

The bodies of the dinosaurs, along with those of other organisms, were deposited in the solid earth and covered by sediment. They might well have simply rotted, and indeed many of them probably did. But many of these organisms were deposited in an anaerobic, or non-oxygen-containing, environment. Rather than simply rotting, this organic material underwent transformation into hydrocarbons and became the basis for the fossil fuels, the most important of which—from the standpoint of modern society—is petroleum. (See Economic Geology for more on this subject.)

CARBONATES

Carbonates are important forms of inorganic carbon in the geosphere. In chemical terms, a carbonate is made from a single carbon atom bonded to three oxygen atoms, but in mineralogical terms, carbonates are a class of mineral that may contain carbon, nitrogen, or boron in a characteristic molecular formation. Typically, a carbonate is transparent and light in color with a relatively high density.



CALCIUM CARBONATE, ONE OF THE MOST COMMON COMPOUNDS IN THE GEOSPHERE, IS FOUND IN SEASHELLS, EGGSHELLS, PEARLS, AND CORAL (PICTURED HERE), BRIDGING THE BOUNDARY BETWEEN THE LIVING AND THE NONLIVING. *(© Fred McConnaughey/Photo Researchers. Reproduced by permission.)*

Among carbonate minerals, the most significant compound is calcium carbonate ($CaCO_3$). One of the most common compounds in the entire geosphere, constituting 7% of the known crustal mass, it is found in such rocks as limestone, marble, and chalk. (Just as pencil "lead" is not really lead, the "chalk" used for writing on blackboards is actually gypsum, a form of calcium sulfate.) Additionally, calcium carbonate can combine with magnesium to form dolomite, and in caves it is the material that makes up stalactites and stalagmites. Yet calcium carbonate also is found in coral, seashells, eggshells, and pearls. This is a good example of how a substance can cross the chemical boundary between the worlds of the living and nonliving.

In the oceans, calcium reacts with dissolved carbon dioxide, forming calcium carbonate and sinking to the bottom. Millions of years ago, when oceans covered much of the planet, sea creatures absorbed calcium and carbon dioxide from the water, which reacted to form calcium carbonate that went into their shells and skeletons. After they died, their bodies became sedimented in the ocean floor, forming vast deposits of limestone.

CARBON DIOXIDE AND CARBON MONOXIDE

Historically, carbon dioxide was the first gas to be distinguished from ordinary air, when in 1630 the Flemish chemist and physicist Jan Baptista van Helmont (1577?–1644) discovered that air was not a single element, as had been thought up to that time. The name perhaps most closely associated with carbon dioxide, however, is that of the English chemist Joseph Priestley (1733–1804), who created carbonated water, used today in making soft drinks. Not only does the gas add bubbles to drinks, it also acts as a preservative.

By Priestley's era, chemists had begun to glimpse a relationship between plant life and carbon dioxide. Up until that time, it had been believed that plants purify the air by day and poison it at night. Today we know that carbon dioxide is an essential component in the natural balance between plant and animal life. Animals, including humans, breathe in air, and, as a result of a chemical reaction in their bodies, the oxygen molecules (O_2) bond with carbon to produce carbon dioxide. Plants "breathe" in this carbon dioxide (which is as important to their survival as air is to animals), and a reverse reaction leads

KEY TERMS

AMORPHOUS: A term for a type of solid that lacks a definite shape. Compare with *crystalline.*

ATOMIC NUMBER: The number of protons in the nucleus of an atom. Since this number is different for each element, elements are listed on the periodic table in order of atomic number.

BIOGEOCHEMICAL CYCLES: The changes that particular elements undergo as they pass back and forth through the various earth systems and particularly between living and nonliving matter. The elements involved in biogeochemical cycles are hydrogen, oxygen, carbon, nitrogen, phosphorus, and sulfur.

CELLULAR RESPIRATION: A process that, when it takes place in the presence of oxygen, involves the intake of organic substances, which are broken down into carbon dioxide and water, with the release of considerable energy.

CHEMICAL BONDING: The joining through electromagnetic force of atoms that sometimes, but not always, represent more than one chemical element. The result is the formation of a molecule.

COMPOUND: A substance made up of atoms of more than one element, chemically bonded to one another.

CRYSTALLINE SOLID: A type of solid in which the constituent parts have a simple and definite geometric arrangement that is repeated in all directions.

DECOMPOSERS: Organisms that obtain their energy from the chemical breakdown of dead organisms as well as from animal and plant waste products. The principal forms of decomposer are bacteria and fungi.

DECOMPOSITION REACTION: A chemical reaction in which a compound is broken down into simpler compounds, or into its constituent elements. In the earth system, this often is achieved through the help of detritivores and decomposers.

ECOSYSTEM: A term referring to a community of interdependent organisms along with the inorganic components of their environment.

to the release of oxygen from the plants back into the atmosphere.

CARBON MONOXIDE. Priestley discovered another carbon-oxygen compound quite different from carbon dioxide: carbon monoxide. The latter is used today by industry for several purposes, such as the production of certain fuels, proving that this toxic gas can be quite beneficial when used in a controlled environment. Nonetheless, carbon monoxide produced in an uncontrolled environment—generated by the burning of petroleum in automobiles as well as by the combustion of wood, coal, and other carbon-containing fuels—is extremely hazardous to human health.

When humans ingest carbon monoxide, it bonds with iron in hemoglobin, the substance in red blood cells that transports oxygen throughout the body. In effect, carbon monoxide fools the body into thinking that it is receiving oxygenated hemoglobin, or oxyhemoglobin. Upon reaching the cells, carbon monoxide has much less tendency than oxygen to break down, and therefore it continues to circulate throughout the body. Low concentrations can cause nausea, vomiting, and other effects, while prolonged exposure to high concentrations can result in death.

THE GREENHOUSE EFFECT. Although we have referred to carbon monoxide

KEY TERMS CONTINUED

ELECTRON: A negatively charged particle in an atom, which spins around the nucleus.

ELEMENT: A substance made up of only one kind of atom. Unlike compounds, elements cannot be broken chemically into other substances.

FOSSIL FUELS: Fuel derived from deposits of organic material that have experienced decomposition and chemical alteration under conditions of high pressure. These nonrenewable forms of bioenergy include petroleum, coal, peat, natural gas, and their derivatives.

GEOCHEMISTRY: A branch of the earth sciences, combining aspects of geology and chemistry, that is concerned with the chemical properties and processes of Earth—in particular, the abundance and interaction of chemical elements and their isotopes.

HYDROCARBON: Any organic chemical compound whose molecules are made up of nothing but carbon and hydrogen atoms.

ORGANIC: At one time chemists used the term *organic* only in reference to living things. Now the word is applied to most compounds containing carbon and hydrogen, thus excluding carbonates (which are minerals) and oxides such as carbon dioxide.

PERIODIC TABLE OF ELEMENTS: A chart that shows the elements arranged in order of atomic number along with the chemical symbol and the average atomic mass for each particular element.

PHOTOSYNTHESIS: The biological conversion of light energy (that is, electromagnetic energy) from the Sun to chemical energy in plants.

PROTON: A positively charged particle in an atom.

REACTIVITY: A term referring to the ability of one element to bond with others. The higher the reactivity, the greater the tendency to bond.

VALENCE ELECTRONS: Electrons that occupy the highest principal energy level in an atom. These are the electrons involved in chemical bonding.

as toxic, it should be noted that carbon dioxide also would be toxic to a human or other animal—for instance, if one were trapped in a sealed compartment and forced to breathe in the carbon dioxide released from one's lungs. On a global scale, both carbon dioxide and carbon monoxide in the atmosphere, produced in excessive amounts by the burning of fossil fuels, pose a potentially serious threat.

Both gases are believed to contribute to the greenhouse effect, which, as discussed in Energy and Earth, is a mechanism by which the planet efficiently uses the heat it receives from the Sun. Human consumption of fossil fuels and use of other products, including chlorofluorocarbons in aerosol cans, however, has produced a much greater quantity of greenhouse gases than the atmosphere needs to maintain normal heat levels. As a result, some scientists believe, buildup of greenhouse gases in the atmosphere is causing global warming.

CELLULAR RESPIRATION

The burning of fossil fuels is one of three ways that carbon enters the atmosphere, the others being volcanic eruption and cellular respiration. When cellular respiration takes place in the presence of oxygen, there is an intake of organic substances, which are broken down into carbon dioxide and water, with the release of considerable energy.

When plants take in carbon dioxide from the atmosphere, they combine it with water and manufacture organic compounds, using energy they have trapped from sunlight by means of photosynthesis—the conversion of light to chemical energy through biological means. As a byproduct of photosynthesis, plants release oxygen into the atmosphere, as we have noted earlier.

In the process of photosynthesis, plants produce carbohydrates, which are various compounds of carbon, hydrogen, and oxygen that are essential to life. (The other two fundamental components of a diet are fats and proteins, both of which are carbon-based as well.) Animals eat the plants or eat other animals that eat the plants and thus incorporate the fats, proteins, and sugars (a form of carbohydrate) from the plants into their bodies. In cellular respiration, these nutrients are broken down to create carbon dioxide.

DECOMPOSITION. Cellular respiration also releases carbon into the atmosphere through the action of decomposers, organisms that obtain their energy from the chemical breakdown of dead organisms as well as from animal and plant waste products. Bacteria and fungi, the principal forms of decomposer, extract energy contained in the chemical bonds of the organic matter they are decomposing and, in the process, release carbon dioxide.

Certain ecosystems, or communities of interdependent organisms, are better than others at producing carbon dioxide through decomposition. As one would expect, environments where heat and moisture are greatest—for example, a tropical rainforest—yield the fastest rates of decomposition. On the other hand, decomposition proceeds much more slowly in dry, cold climates such as that of a subarctic tundra.

WHERE TO LEARN MORE

Blashfield, Jean F. *Carbon.* Austin, TX: Raintree Steck–Vaughn, 1999.

Carbon Cycle: Exploring the Environment (Web site). <http://www.cotf.edu/ete/modules/carbon/efcarbon.html>.

Carbon Cycle Greenhouse Gases Group (CCGG) (Web site). <http://www.cmdl.noaa.gov/ccgg/>.

Chemical Carousel: A Trip Around the Carbon Cycle (Web site). <http://library.thinkquest.org/11226/>.

Frostburg State University Chemistry Helper (Web site). <http://www.chemhelper.com/>.

"Global Carbon Cycle." The Woods Hole Research Center (Web site). <http://www.whrc.org/science/carbon/carbon.htm>.

Knapp, Brian J. *Carbon Chemistry.* Illus. David Woodroffe. Danbury, CT: Grolier Educational, 1998.

Loudon, G. Marc. *Organic Chemistry.* Menlo Park, CA: Benjamin/Cummings, 1988.

Sparrow, Giles. *Carbon.* New York: Benchmark Books, 1999.

Stille, Darlene. *The Respiratory System.* New York: Children's Press, 1997.

THE NITROGEN CYCLE

CONCEPT

Contrary to popular belief, the air we breathe is not primarily oxygen; by far the greatest portion of air is composed of nitrogen. A colorless, odorless gas noted for its lack of chemical reactivity—that is, its tendency not to bond with other elements—nitrogen plays a highly significant role within the earth system. Both through the action of lightning in the sky and of bacteria in the soil, nitrogen is converted to nitrites and nitrates, compounds of nitrogen and oxygen that are then absorbed by plants to form plant proteins. The latter convert to animal proteins in the bodies of animals who eat the plants, and when an animal dies, the proteins are returned to the soil. Denitrifying bacteria break down these compounds, returning elemental nitrogen to the atmosphere.

HOW IT WORKS

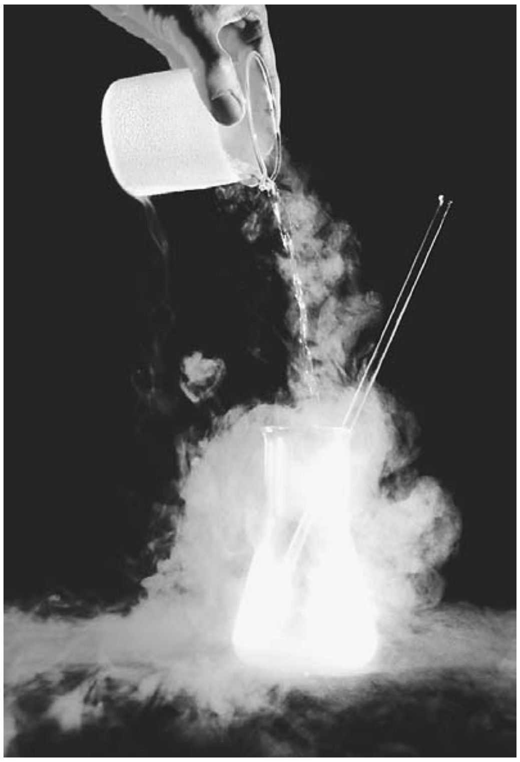
Chemistry and Elements

The concepts we discuss in this essay fall under the larger heading of geochemistry. A branch of the earth sciences that combines aspects of geology and chemistry, geochemistry is concerned with the chemical properties and processes of Earth. Among particular areas of interest in geochemistry are biogeochemical cycles, or the changes that particular elements undergo as they pass back and forth through the various earth systems and particularly between living and nonliving matter. The elements involved in biogeochemical cycles are hydrogen, oxygen, carbon, nitrogen, phosphorus, and sulfur (see Biogeochemical Cycles and Carbon Cycle).

An element is a substance composed of a single type of atom, which cannot be broken down chemically into a simpler substance. Each element is distinguished by its atomic number, or the number of protons (positively charged subatomic particles) in the nucleus, or center, of the atom. On the periodic table of elements, these fundamental substances of the universe are listed in order of atomic number, from hydrogen to uranium—which has the highest atomic number (92) of any element that occurs in nature—and beyond. The elements with an atomic number higher than that of uranium, all of which have been created artificially, play virtually no role in the chemical environment of Earth and are primarily of interest only to specialists in certain fields of chemistry and physics.

Everything that exists in the universe is an element, a compound formed by the chemical bonding of elements, or a mixture of compounds. In order to bond and form a compound, elements experience chemical reactions, which are the result of attractions on the part of electrons (negatively charged subatomic particles) that occupy the highest energy levels in the atom. These electrons are known as valence electrons.

CHEMICAL CHANGES. A chemical change is a phenomenon quite different from a physical change. If liquid water boils or freezes (both of which are examples of a physical change resulting from physical processes), it is still water. Physical changes do not affect the internal composition of an item or items; a chemical change, on the other hand, occurs when the actual composition changes—that is, when one substance is transformed into another. Chemical change requires a chemical reaction, a process whereby



LIQUID NITROGEN. *(© David Taylor/Photo Researchers. Reproduced by permission.)*

the chemical properties of a substance are altered by a rearrangement of atoms.

There are several clues that tell us when a chemical reaction has taken place. In many chemical reactions, for instance, the substance may experience a change of state or phase—as, for instance, when liquid water is subjected to an electric current through a process known as *electrolysis,* which separates it into oxygen and hydrogen, both of which are gases. Another clue that a chemical reaction has occurred is a change of temperature. Unlike the physical change of liquid water to ice or steam, however, this temperature change involves an alteration of the chemical properties of the substances themselves. Chemical reactions also may encompass changes in color, taste, or smell.

Nitrogen's Place Among the Elements

With an atomic number of 7, nitrogen (chemical symbol N) is one of just 19 elements that are nonmetals. Unlike metals, nonmetals are poor conductors of heat and electricity and are not ductile—in other words, they cannot be reshaped easily. The vast majority of elements are metallic, however, the only exceptions being the nonmetals as well as six "metalloids," or elements that display characteristics of both metals and nonmetals.

Nitrogen is also one of eight "orphan" nonmetals—those nonmetals that do not belong to any family of elements, such as the halogens or noble gases. All six of the elements involved in biogeochemical cycles, in fact, are "orphan" nonmetals, with boron and selenium rounding out the list of eight orphans. Sometimes nitrogen is considered the head of a "family" of elements, all of which occupy a column or group on the periodic table.

These five elements—nitrogen, phosphorus, arsenic, antimony, and bismuth—share a common pattern of valence electrons, but otherwise they share little in terms of physical properties or chemical behavior. By contrast, chemicals that truly are related all have a common "family resemblance": all halogens are highly reactive, for instance, while all noble gases are extremely unreactive.

ABUNDANCE. The seventeenth most abundant element on Earth, nitrogen accounts for 0.03% of the planet's known elemental mass. This may seem very small, but at least nitrogen is among the 18 elements considered relatively abundant. These 18 elements account for all but 0.49% of the planet's known elemental mass, the remainder being composed of numerous other elements in small quantities. The term known elemental mass takes account of the fact that scientists do not know with certainty the elemental composition of Earth's interior, though it likely contains large proportions of iron and nickel. The known mass, therefore, is that which exists from the bottom of the crust to the top layers of the atmosphere.

Elemental proportions too small to be measured in percentage points are rendered in parts per million (ppm) or even parts per billion (ppb). Within the crust itself, nitrogen's share is certainly modest: a concentration of 19 ppm, which ties it with gallium, a metal whose name is hardly a household word, for a rank of thirty-third. On the other hand, this still makes it more abundant in the crust than many quite familiar metals, including lithium, uranium, tungsten, silver, mercury, and platinum.

In Earth's atmosphere, on the other hand, the proportion of nitrogen is much, much higher. The atmosphere is 78% nitrogen and 21% oxygen, while the noble gas argon accounts for 0.93%. The remaining 0.07% is taken up by various trace gases, including water vapor, carbon dioxide, and ozone, or O_3.

In the human body, nitrogen's share is much more modest than it is in the atmosphere but still 10 times greater than it is in relation to the planet's total mass. The element accounts for 3% of the body's mass, making it the fourth most abundant element in the human organism.

Properties and Applications of Nitrogen

The Scottish chemist Daniel Rutherford (1749–1819) usually is given credit for discovering nitrogen in 1772, when he identified it as the element that remained when oxygen was removed from air. Several other scientists at about the same time made a similar discovery.

Because of its heavy presence in air, nitrogen is obtained primarily by cooling air to temperatures below the boiling points of its major components. Nitrogen boils (that is, turns into a gas) at a lower temperature than oxygen: –320.44°F (–195.8°C), as opposed to –297.4°F (–183°C). If air is cooled to –328°F (–200°C), thus solidifying it, and then allowed to warm slowly, the nitrogen boils first and therefore evaporates first. The nitrogen gas is captured, cooled, and liquefied once more.

Nitrogen also can be obtained from such compounds as potassium nitrate or saltpeter, found primarily in India, or from sodium nitrate (Chilean saltpeter), which comes from the desert regions of Chile. To isolate nitrogen chemically, various processes are undertaken in a laboratory—for instance, heating barium azide or sodium azide, both of which contain nitrogen.

REACTIONS WITH OTHER ELEMENTS. Rather than appearing as single atoms, nitrogen is diatomic, meaning that two nitrogen atoms typically bond with each other to form dinitrogen, or N_2. Nor do these atoms form single chemical bonds, as is characteristic of most elements; theirs is a *triple* bond, which effectively ties up the atoms' valence electrons, making nitrogen an unreactive element at relatively low temperatures.

Even at the temperature of combustion, a burning substance reacts with the oxygen in the air but not with the nitrogen. At very high temperatures, on the other hand, nitrogen combines with other elements, reacting with metals to form nitrides, with hydrogen to form ammonia, with O_2 (oxygen as it usually appears in nature, two atoms bonded in a molecule) to form nitrites, and with O_3 (ozone) to form nitrates. With the exception of the first-named group, all of these elements are important to our discussion of nitrogen.

Nitrogen combines with hydrogen to form ammonia. Ammonium nitrate, a fertilizer, is also a dangerous explosive. It was used in April of 1995 to blow up the Alfred P. Murrah Federal building in Oklahoma City, killing 168 people. *(© James H. Robinson/Photo Researchers. Reproduced by permission.)*

SOME USES FOR NITROGEN. In processing iron or steel, which forms undesirable oxides if exposed to oxygen, a blanket of nitrogen is applied to prevent this reaction. The same principle is applied in making computer chips and even in processing foods, since these items, too, are affected detrimentally by oxidation. Because it is far less combustible than air (magnesium is one of the few elements that burns nitrogen in combustion), nitrogen also is used to clean tanks that have carried petroleum or other combustible materials.

As noted, nitrogen combines with hydrogen to form ammonia, used in fertilizers and cleaning materials. Ammonium nitrate, applied primarily as a fertilizer, is also a dangerous explosive, as shown with horrifying effect in the bombing of the Alfred P. Murrah Federal Building in Oklahoma City on April 19, 1995—a tragedy that took 168 lives. Nor is ammonium nitrate the only nitrogen-based explosive. Nitric acid is used in making trinitrotoluene (TNT), nitroglycerin, and dynamite as well as gunpowder and smokeless powder.

INTRODUCTION TO THE NITROGEN CYCLE

The nitrogen cycle is the process whereby nitrogen passes from the atmosphere into living things and ultimately back into the atmosphere. In the process, it is converted to nitrates and nitrites, compounds of nitrogen and oxygen that are absorbed by plants in the process of forming plant proteins. These plant proteins, in turn, are converted to animal proteins in the bodies of animals who eat the plants, and when the animal dies, the proteins are returned to the soil. Denitrifying bacteria break down these organic compounds, returning elemental nitrogen to the atmosphere.

Note what happens in the nitrogen cycle and, indeed, in all biogeochemical cycles: organic material is converted to inorganic material through various processes, and inorganic material absorbed by living organisms eventually is turned into organic material. In effect, the element passes back and forth between the realms of the living and the nonliving. This may sound a bit mystical, but it is not. To be organic, a substance must be built around carbon in certain characteristic chemical structures, and by inducing the proper chemical reaction, it is possible to break down or build up these structures, thus turning an organic substance into an inorganic one, or vice versa. (For more on this subject, see Carbon Cycle.)

STEPS IN THE CYCLE. Plants depend on biologically useful forms of nitrogen, the availability of which greatly affects their health, abundance, and productivity. This is particularly the case where plants in a saltwater ecosystem (a community of interdependent organisms) are concerned. Regardless of the specific ecosystem, however, fertilization of the soil with nitrogen has an enormous impact on the growth yield of plant life, which can be critical in the case of crops. Therefore, nitrogen is by far the most commonly applied nutrient in an agricultural setting.

There are several means by which plants receive nitrogen. They may absorb it as nitrate or ammonium, dissolved in saltwater and taken up through the roots, or as various nitrogen oxide gases. In certain situations, plants have a symbiotic, or mutually beneficial, relationship with microorganisms capable of "fixing" atmospheric dinitrogen into ammonia. In any case, plants receive nitrogen and later, when they are eaten by animals, pass these nutrients along the food chain—or rather, to use a term more favored in the earth and biological sciences, the food web.

When herbivorous or omnivorous animals consume nitrogen-containing plants, their bodies take in the nitrogen and metabolize it, breaking it down to generate biochemicals, or chemicals essential to life processes. At some point, the animal dies, and its body experiences decomposition through the activity of bacteria and other decomposers. These microorganisms, along with detritivores such as earthworms, convert nitrates and nitrites from organic sources into elemental nitrogen, which ultimately reenters the atmosphere.

REAL-LIFE APPLICATIONS

IMPORTANT FORMS OF NITROGEN

As noted earlier, dinitrogen, or N_2, is the form in which nitrogen typically appears when uncombined with other elements. This is also the form of nitrogen in the atmosphere, but it is so chemically unreactive that unlike oxygen, it plays little actual part in sustaining life. Indeed, because nitrogen in the air is essentially "filler" as far as humans are concerned, it can be substituted with helium, as is done in air tanks for divers. This prevents them from experiencing decompression sickness, or "the bends," which occurs when the diver returns too quickly to the surface, causing nitrogen in the blood to boil.

The dinitrogen in the air is a holdover from long ago in Earth's development, when volcanoes expelled elements from deep in the planet's interior to its atmosphere. Owing to its lack of reac-

SMOG BLANKETS LOS ANGELES IN A HAZE. NITRIC OXIDE REACTS WITH OXYGEN IN THE AIR TO FORM NITROGEN OXIDE, A REDDISH BROWN GAS THAT COLORS SMOG. *(Photograph by Walter A. Lyons. FMA Productions. Reproduced by permission.)*

tivity, dinitrogen never went anywhere. For it to play a role in the functioning of Earth cycles, it must be "fixed," as discussed later in this essay. In addition to dinitrogen, nitrogen appears in a number of other important inorganic compounds, including nitrite and nitrate; ammonia and ammonium; and nitric oxide, nitrogen dioxide, and nitrous oxide.

Nitrite and nitrate are two ionic forms of nitrogen. An ion is an atom or group of atoms that has lost or gained electrons, thus acquiring a net electric charge. Both nitrite and nitrate are anions, or negatively charged ions, designated by the use of superscript minus signs that indicate that each has a net charge of negative 1. Thus, nitrite, in which nitrogen is chemically bonded with two atoms of oxygen, is rendered as NO_2^-, while the formula for nitrate (nitrogen with three oxygen atoms), is designated as NO_3^-.

AMMONIA AND AMMONIUM. Nitrification is a process in which nitrite is produced, whereupon it undergoes a chemical reaction to form nitrate, the principal form of nitrogen nutrition for most plant species. The chemical from which the nitrite is created in the nitrification reaction is ammonium (NH_4^+), which is formed by the addition of a hydrogen cation, or a positively charged ion (H^+), to ammonia, or NH_3. The latter, which is probably familiar to most people in the form of a household cleaner, is actually an extremely abundant compound, both in natural and artificial forms.

Ammonium is soluble, or capable of being dissolved, in water and often is used as a fertilizer. It is attracted to negatively charged surfaces of clays and organic matter in soil and therefore tends to become stuck in one place rather than moving around, as nitrate does. In acidic soils, typically plants receive their nitrogen from ammonium, but most nonacidic soils can use only nitrate. As noted earlier, ammonium may be combined with nitrate to form ammonium nitrate—both a powerful fertilizer and a powerful explosive.

OXIDES. Nitrogen reenters the atmosphere in the form of the gas nitric oxide (NO), emitted primarily as the result of combustion reactions. This may occur in one of two ways. Organic nitrogen in bioenergy sources, such as biomass (organisms, their waste products, and their incompletely decomposed remains) or fossil fuels (e.g., coal or oil), may be oxidized. The latter term means that a substance undergoes a chemical reaction with oxygen: combustion itself, which requires the presence of oxygen, is an example of oxidation.

KEY TERMS

ATOM: The smallest particle of an element, consisting of protons, neutrons, and electrons. An atom can exist either alone or in combination with other atoms in a molecule.

ATOMIC NUMBER: The number of protons in the nucleus of an atom. Since this number is different for each element, elements are listed on the periodic table in order of atomic number.

BIOENERGY: Energy derived from biological sources that are used directly as fuel (as opposed to food, which becomes fuel).

BIOGEOCHEMICAL CYCLES: The changes that particular elements undergo as they pass back and forth through the various earth systems and specifically between living and nonliving matter. The elements involved in biogeochemical cycles are hydrogen, oxygen, carbon, nitrogen, phosphorus, and sulfur.

CHEMICAL BONDING: The joining, through electromagnetic force, of atoms that sometimes, but not always, represent more than one chemical element. The result is usually the formation of a molecule.

COMPOUND: A substance made up of atoms of more than one element chemically bonded to one another.

DECOMPOSERS: Organisms that obtain their energy from the chemical breakdown of dead organisms as well as from animal and plant waste products. The principal forms of decomposer are bacteria and fungi.

DECOMPOSITION REACTION: A chemical reaction in which a compound is broken down into simpler compounds or into its constituent elements. In the earth system, this often is achieved through the help of detritivores and decomposers.

DETRITIVORES: Organisms that feed on waste matter, breaking organic material down into inorganic substances that then become available to the biosphere in the form of nutrients for plants. Their function is similar to that of decomposers; however, unlike decomposers—which tend to be bacteria or fungi—detritivores are relatively complex organisms, such as earthworms or maggots.

DIATOMIC: A term describing a chemical element that typically exists as molecules composed of two atoms. Nitrogen and oxygen are both diatomic.

ECOSYSTEM: A term referring to a community of interdependent organisms along with the inorganic components of their environment.

On the other hand, nitric oxide may enter the atmosphere when atmospheric dinitrogen is combined with oxygen under conditions of high temperature and pressure, as, for instance, in an internal-combustion engine. In the atmosphere, nitric oxide reacts readily with oxygen in the air to form nitrogen dioxide (NO_2), a reddish-brown gas that adds to the tan color of smog over major cities.

Yet nitric oxide and nitrogen dioxide, usually designated together as NO_x, are also part of the life-preserving nitrogen cycle. Gaseous NO_x is taken in by plants, or oxidized to make nitrate, and circulated through the biosphere or else cycled directly to the atmosphere. In addition, denitrification, discussed later in this essay, transports nitrous oxide (N_2O) into the atmosphere from nitrate-rich soils.

KEY TERMS

ELECTRON: A negatively charged particle in an atom, which spins around the nucleus.

ELEMENT: A substance made up of only one kind of atom. Unlike compounds, elements cannot be chemically broken into other substances.

EUTROPHICATION: A state of heightened biological productivity in a body of water, which is typically detrimental to the ecosystem in which it takes place. Eutrophication can be caused by an excess of nitrogen or phosphorus in the form of nitrates and phosphates, respectively.

FOOD WEB: A term describing the interaction of plants, herbivores, carnivores, omnivores, decomposers, and detritivores, each of which consumes nutrients and passes it along to other organisms.

GEOCHEMISTRY: A branch of the earth sciences combining aspects of geology and chemistry, that is, concerned with the chemical properties and processes of Earth—in particular, the abundance and interaction of chemical elements and their isotopes.

ION: An atom or group of atoms that has lost or gained one or more electrons and thus has a net electric charge. Positively charged ions are called *cations,* and negatively charged ones are called *anions.*

LEACHING: The removal of soil materials that are in solution, or dissolved in water.

MOLECULE: A group of atoms, usually but not always representing more than one element, joined in a structure. Compounds are typically made up of molecules.

PERIODIC TABLE OF ELEMENTS: A chart that shows the elements arranged in order of atomic number along with their chemical symbols and the average atomic mass for each particular element.

PROTON: A positively charged particle in an atom.

REACTIVITY: A term referring to the ability of one element to bond with others. The higher the reactivity, the greater the tendency to bond.

SOLUBLE: Capable of being dissolved.

VALENCE ELECTRONS: Electrons that occupy the highest principal energy level in an atom. These are the electrons involved in chemical bonding.

NITROGEN PROCESSES

In order for most organisms to make use of atmospheric dinitrogen, it must be "fixed" into inorganic forms that a plant can take in through its roots and leaves. Nonbiological processes, such as a lightning strike, can bring about dinitrogen fixation. The high temperatures and pressures associated with lightning lead to the chemical bonding of atmospheric nitrogen and oxygen (both of which appear in diatomic form) to create two molecules of nitric oxide.

More often than not, however, dinitrogen fixation comes about through biological processes. Microorganisms are able to synthesize an enzyme that breaks the triple bonds in dinitrogen, resulting in the formation of two molecules of ammonia for every dinitrogen molecule thus reacted. This effect is achieved most commonly by bacteria or algae in wet or moist environments that offer nutrients other than nitrate or ammonium. In some instances, plants enjoy a symbiotic, or mutually beneficial, relationship with microorganisms capable of fixing dinitrogen.

AMMONIFICATION, NITRIFICATION, AND DENITRIFICATION. Dinitrogen fixation is just one example of a process whereby nitrogen is processed through one or more earth systems. Another is ammonification, or the process whereby nitrogen in organisms is recycled after their death. Enabled by microorganisms that act as decomposers, ammonification results in the production of either ammonia or ammonium. Thus, the soil is fertilized by the decayed matter of formerly living things.

Ammonium, as we noted earlier, also plays a part in nitrification, a process in which it first is oxidized to produce nitrite. Then the nitrite is oxidized to become nitrate, which fertilizes the soil. As previously mentioned, nitrate is useful as a fertilizer only in non-acidic soils; acidic ones, by contrast, require ammonium fertilizer.

In contrast to nitrification is denitrification, in which nitrate is reduced to the form of either nitrous oxide or dinitrogen. This takes place under anaerobic conditions—that is, in the absence of oxygen—and on the largest scale when concentrations of nitrate are highest. Flooded fields, for example, may experience high rates of denitrification.

The Role of Humans

Humans are involved in the nitrogen cycle in several ways, not all of them beneficial. One of the most significant roles people play in the nitrogen cycle is by the introduction of nitrogen-containing fertilizers to the soil. Because nitrogen has a powerful impact on plant growth, farmers are tempted to add more and more nitrate or ammonium or both to their crops, to the point that the soil becomes saturated with it and therefore unable to absorb more.

When the soil has taken in all the nitrogen it can hold, a process of leaching—the removal of soil materials dissolved in water—eventually takes place. Nitrate, in particular, leaches from agricultural sites into groundwater as well as streams and other forms of surface water. This can lead to eutrophication, a state of heightened biological productivity that is ultimately detrimental to the ecosystem surrounding a lake or other body of water. (See Biogeochemical Cycles for more about eutrophication.)

Yet another problem associated with overly nitrate-rich soils is an excessive rate of denitrification. This happens when soils that have been loaded down with nitrates become wet for long periods of time, leading to a dramatic increase in the denitrification rate. As a result, fixed nitrogen is lost, and nitrous oxide is emitted to the air. In the atmosphere nitrous oxide may contribute to the greenhouse effect, possibly helping increase the overall temperature of the planet (see Carbon Cycle and Energy and Earth).

WHERE TO LEARN MORE

Blashfield, Jean F. *Nitrogen.* Austin, TX: Raintree Steck–Vaughn, 1999.

Farndon, John. *Nitrogen.* New York: Benchmark Books, 1999.

Fitzgerald, Karen. *The Story of Nitrogen.* New York: Franklin Watts, 1997.

The Microbial World: The Nitrogen Cycle and Nitrogen Fixation (Web site). <http://helios.bto.ed.ac.uk/bto/microbes/nitrogen.htm>.

The Nitrogen Cycle (Web site). <http://clab.cecil.cc.md.us/faculty/biology/jason/nitrc.htm>.

The Nitrogen Cycle (Web site). <http://www.ultranet.com/~jkimball/BiologyPages/N/NitrogenCycle.html>.

The Nitrogen Cycle (Web site). <http://library.thinkquest.org/11353/nitrogen.htm>.

Nutrient Overload: Unbalancing the Global Nitrogen Cycle (Web site). <http://www.wri.org/wri/wr-98-99/nutrient.htm>.

Postgate, J. R. *The Outer Reaches of Life.* New York: Cambridge University Press, 1994.

Skinner, Brian J., Stephen C. Porter, and Daniel B. Botkin. *The Blue Planet: An Introduction to Earth System Science.* 2d ed. New York: John Wiley and Sons, 1999.

SCIENCE OF EVERYDAY THINGS
REAL-LIFE EARTH SCIENCE

THE BIOSPHERE

ECOSYSTEMS

CONCEPT

An ecosystem is a complete community of living organisms and the nonliving materials of their surroundings. Thus, its components include plants, animals, and microorganisms; soil, rocks, and minerals; as well as surrounding water sources and the local atmosphere. The size of ecosystems varies tremendously. An ecosystem could be an entire rain forest, covering a geographical area larger than many nations, or it could be a puddle or a backyard garden. Even the body of an animal could be considered an ecosystem, since it is home to numerous microorganisms. On a much larger scale, the history of various human societies provides an instructive illustration as to the ways that ecosystems have influenced civilizations.

HOW IT WORKS

The Biosphere

Earth itself could be considered a massive ecosystem, in which the living and nonliving worlds interact through four major subsystems: the atmosphere, hydrosphere (all the planet's waters, except for moisture in the atmosphere), geosphere (the soil and the extreme upper portion of the continental crust), and biosphere. The biosphere includes all living things: plants (from algae and lichen to shrubs and trees), mammals, birds, reptiles, amphibians, aquatic life, insects, and all manner of microscopic forms, including bacteria and viruses. In addition, the biosphere draws together all formerly living things that have not yet decomposed.

Several characteristics unite the biosphere. One is the obvious fact that everything in it is either living or recently living. Then there are the food webs that connect organisms on the basis of energy flow from one species to another. A food web is similar to the more familiar concept food chain, but in scientific terms a food chain—a series of singular organisms in which each plant or animal depends on the organism that precedes or follows it—does not exist. Instead, the feeding relationships between organisms in the real world are much more complex and are best described as a web rather than a chain.

Food Webs

Food webs are built around the flow of energy between organisms, known as energy transfer, which begins with plant life. Plants absorb energy in two ways. From the Sun, they receive electromagnetic energy in the form of visible light and invisible infrared waves, which they convert to chemical energy through a process known as photosynthesis. In addition, plants take in nutrients from the soil, which contain energy in the forms of various chemical compounds. These compounds may be organic, which typically means that they came from living things, though, in fact, the term *organic* refers strictly to characteristic carbon-based chemical structures. Plants also receive inorganic compounds from minerals in the soil. (See Minerals. For more about the role of carbon in inorganic compounds, see Carbon Cycle.)

Contained in these minerals are six chemical elements essential to the sustenance of life on planet Earth: hydrogen, oxygen, carbon, nitrogen, phosphorus, and sulfur. These are the elements involved in biogeochemical cycles, through which they continually are circulated between the living and nonliving worlds—that is, between organisms, on the one hand, and the inorganic realms of rocks, minerals, water, and air, on the other (see Biogeochemical Cycles).

FROM PLANTS TO CARNIVORES. As plants take up nutrients from the soil, they convert them into other forms, which provide usable energy to organisms who eat the plants. (An example of this conversion process is cellular respiration, discussed in Carbon Cycle.) When an herbivore, or plant-eating organism, eats the plant, it incorporates this energy.

Chances are strong that the herbivore will be eaten either by a carnivore, a meat-eating organism, or by an omnivore, an organism that consumes both herbs and herbivores—that is, both plants and animals. Few animals consume carnivores or omnivores, at least by hunting and killing them. (Detritivores and decomposers, which we discuss presently, consume the remains of all creatures, including carnivores and omnivores.) Humans are an example of omnivores, but they are far from the only omnivorous creatures. Many bird species, for instance, are omnivorous.

As nutrients pass from plant to herbivore to carnivore, the total amount of energy in them decreases. This is dictated by the second law of thermodynamics (see Energy and Earth), which shows that energy transfers cannot be perfectly efficient. Energy is not "lost"—the total amount of energy in the universe remains fixed, though it may vary with a particular system, such as an individual ecosystem—but it is dissipated, or directed into areas that do not aid in the transfer of energy between organisms. What this means for the food web is that each successive level contains less energy than the levels that precede it.

DETRITIVORES AND DECOMPOSERS. In the case of a food web, something interesting happens with regard to energy efficiency as soon as we pass beyond carnivores and omnivores to the next level. It might seem at first that there could be no level beyond carnivores or omnivores, since they appear to be "at the top of the food chain," but this only illustrates why the idea of a food web is much more useful. After carnivores and omnivores, which include some of the largest, most powerful, and most intelligent creatures, come the lowliest of all organisms: decomposers and detritivores, an integral part of the food web.

Decomposers, which include bacteria and fungi, obtain their energy from the chemical breakdown of dead organisms as well as from animal and plant waste products. Detritivores perform a similar function: by feeding on waste matter, they break organic material down into inorganic substances that then can become available to the biosphere in the form of nutrients for plants. The principal difference between detritivores and decomposers is that the former are relatively complex organisms, such as earthworms or maggots.

Both decomposers and detritivores aid in decomposition, a chemical reaction in which a compound is broken down into simpler compounds or into its constituent elements. Often an element such as nitrogen appears in forms that are not readily usable by organisms, and therefore such elements (which may appear individually or in compounds) need to be chemically processed through the body of a decomposer or detritivore. This processing involves chemical reactions in which the substance—whether an element or compound—is transformed into a more usable version.

By processing chemical compounds from the air, water, and geosphere, decomposers and detritivores deposit nutrients in the soil. These creatures feed on plant life, thus making possible the cycle we have described. Clearly this system, of which we have sketched only the most basic outlines, is an extraordinarily complex and well-organized one, in which every organism plays a specific role. In fact, earth scientists working in the realm of biosphere studies use the term niche to describe the role that a particular organism plays in its community. (For more about the interaction of species in a biological community, see Ecology and Ecological Stress.)

REAL-LIFE APPLICATIONS

The Fate of Human Civilizations

An interesting place to start in investigating examples of ecosystems is with a species near and dear to all of us: *Homo sapiens.* Much has been written about the negative effect industrial civilization has, or may have, on the natural environment—a topic discussed in Ecology and Ecological Stress—but here our concern is somewhat different. What do ecosystems, and specifically the availability of certain plants and animals, teach us about specific societies?

In his 1997 bestseller *Guns, Germs, and Steel: The Fate of Human Societies,* the ethnobotanist Jared M. Diamond (1937-) explained how he came to approach this question. While he was working with native peoples in New Guinea, a young man asked him why the societies of the West enjoyed an abundance of material wealth and comforts while those of New Guinea had so little. It was a simple question, but the answer was not obvious.

Diamond refused to give any of the usual pat responses offered in the past—for example, the Marxist or socialist claim that the West prospers at the expense of native peoples. Nor, of course, could he accept the standard answer that a white descendant of Europeans might have given a century earlier, that white Westerners are smarter than dark-skinned peoples. Instead, he approached it as a question of environment, and the result was his thought-provoking analysis contained in *Guns, Germs, and Steel.*

ADVANTAGES OF GEOGRAPHY. As Diamond showed, those places where agriculture was first born were precisely those blessed with favorable climate, soil, and indigenous plant and animal life. Incidentally, none of these locales was European, nor were any of the peoples inhabiting them "white." Agriculture came into existence in four places during a period from about 8000 to 6000 B.C. In roughly chronological order, they were Mesopotamia, Egypt, India, and China. All were destined to emerge as civilizations, complete with written language, cities, and organized governments, between about 3000 and 2000 B.C.

Of course, it is no accident that civilization was born first in those societies that first developed agriculture: before a civilization can evolve, a society must become settled, and in order for that to happen, it must develop agriculture. Each of these societies, it should be noted, formed along a river, and that of Mesopotamia was born at the confluence of *two* rivers, the Tigris and Euphrates. No wonder, then, that the spot where these two rivers met was identified in the Bible as the site for the Garden of Eden or that historians today refer to ancient Mesopotamia as "the Fertile Crescent." (For a very brief analysis regarding possible reasons why modern Mesopotamia—that is, Iraq—does not fit this description, see the discussion of desertification in Soil Conservation.)

In the New World, by contrast, agriculture appeared much later and in a much more circumscribed way. The same was true of Africa and the Pacific Islands. In seeking the reasons for why this happened, Diamond noted a number of factors, including geography. The agricultural areas of the Old World were stretched across a wide area at similar latitudes. This meant that the climates were not significantly different and would support agricultural exchanges, such as the spread of wheat and other crops from one region or ecosystem to another. By contrast, the land masses of the New World or Africa have a much greater north-south distance than they do east to west.

DIVERSITY OF SPECIES. Today such places as the American Midwest support abundant agriculture, and one might wonder why that was not the case in the centuries before Europeans arrived. The reason is simple but subtle, and it has nothing to do with Europeans' "superiority" over Native Americans. The fact is that the native North American ecosystems enjoyed far less biological diversity, or biodiversity, than their counterparts in the Old World. Peoples of the New World successfully domesticated corn and potatoes, because those were available to them. But they could not domesticate emmer wheat, the variety used for making bread, when they had no access to that species, which originated in Mesopotamia and spread throughout the Old World.

Similarly, the New World possessed few animals that could be domesticated either for food or labor. A number of Indian tribes domesticated some types of birds and other creatures for food, but the only animal ever adapted for labor was the llama. The llama, a cousin of the camel found in South America, is too small to carry heavy loads. Why did the Native Americans never harness the power of cows, oxen, or horses? For the simple reason that these species were not found in the Americas. After horses in the New World went extinct at some point during the last Ice Age (see Paleontology), they did not reappear in the

THE LLAMA WAS ONE OF THE FEW DOMESTICATED ANIMALS ADAPTED FOR WORK IN THE NEW WORLD, A PLACE WITH A SMALL NUMBER OF ANIMAL AND PLANT SPECIES AND LACK OF ECOLOGICAL COMPLEXITY BEFORE THE EUROPEANS ARRIVED IN ABOUT 1500 A.D. *(© Francois Gohier/Photo Researchers. Reproduced by permission.)*

Americas until Europeans brought them after A.D. 1500.

Diamond also noted the link between biodiversity and the practice, common among peoples in New Guinea and other remote parts of the world, of eating what Westerners would consider strange cuisine: caterpillars, insects—even, in some cases, human flesh. At one time, such practices served only to brand these native peoples further as "savages" in the eyes of Europeans and their descendants, but it turns out that there is a method to the apparent madness. In places such as the highlands of New Guinea, a scarcity of animal protein sources compels people to seek protein wherever they can find it.

By contrast, from ancient times the Fertile Crescent possessed an extraordinary diversity of animal life. Among the creatures present in that region (the term sometimes is used to include Egypt as well as Mesopotamia) were sheep, goats, cattle, pigs, and horses. With the help of these animals for both food and labor—people ate horses long before they discovered their greater value as a mode of transportation—the lands of the Old World were in a position to progress far beyond their counterparts in the New.

GREATER EXPOSURE TO MICROORGANISMS. Ultimately, these societies came to dominate their physical environments and excel in the development of technology; hence the "steel" and "guns" in Diamond's title. But what about "germs"? It is a fact that after Europeans began arriving in the New World, they killed vast populations without firing a shot, thanks to the microbes they carried with them. Of course, it would be centuries before scientists discovered the existence of microorganisms. But even in 1500, it was clear that the native peoples of the New World had no natural resistance to smallpox or a host of other diseases, including measles, chicken pox, influenza, typhoid fever, and bubonic plague.

Once again the Europeans' advantage over the Native Americans derived from the ecological complexity of their world compared with that of the Indians. In the Old World, close contact with farm animals exposed humans to germs and disease. So, too, did close contact with other people in crowded, filthy cities. This exposure, of course, killed off large numbers of people, but those who survived tended to be much hardier and possessed much stronger immune systems. Therefore, when Europeans came into contact with

With its sauna-like environment and closed canopy, a tropical cloud forest produces lush vegetation and is one of the most biodiverse ecosystems on the earth. *(© G. Dimijian/Photo Researchers. Reproduced by permission.)*

native Americans, they were like walking biological warfare weapons.

Evaluating Ecosystems

The ease with which Europeans subdued Native Americans fueled the belief that Europeans were superior, but, as Diamond showed, if anything was superior, it was the ecosystems of the Old World. This "superiority" relates in large part to the diversity of organisms an ecosystem possesses. Many millions of years ago, Earth's oceans and lands were populated with just a few varieties of single-cell organisms, but over time increasing differentiation of species led to the development of the much more complex ecosystems we know now.

Such differentiation is essential, given the many basic types of ecosystem that the world has to offer: forests and grasslands, deserts and aquatic environments, mountains and jungles. Among the many ways that these ecosystems can be evaluated, aside from such obvious parameters as relative climate, is in terms of abundance and complexity of species.

ABUNDANCE AND COMPLEXITY. The biota (a combination of all flora and fauna, or plant and animal life, respectively) in a desert or the Arctic tundra is much less complex than that of a tropical rain forest or, indeed, almost any kind of forest, because far fewer species can live in a desert or tundra environment. For this reason, it is said that a desert or tundra ecosystem is less complex than a forest one. There may be relatively large numbers of particular species in a less complex ecosystem, however, in which case the ecosystem is said to be abundant though not complex in a relative sense.

Another way to evaluate ecosystems is in terms of productivity. This concept refers to the amount of biomass—potentially burnable energy—produced by green plants as they capture sunlight and use its energy to create new organic compounds that can be consumed by local animal life. Once again, a forest, and particularly a rain forest, has a very high level of productivity, whereas a desert or tundra ecosystem does not.

Forests

Now let us look more closely at a full-fledged ecosystem—that of a forest—in action. It might seem that all forests are the same, but this could not be less the case. A forest is simply any ecosystem dominated by tree-sized woody plants. Beyond that, the characteristics of weather, climate, elevation, latitude, topography, tree species, varieties of animal species, moisture levels, and numerous other parameters create the potential for an almost endless diversity of forest types.

In fact, the United Nations Educational, Scientific, and Cultural Organization (UNESCO) defines 24 different types of forest, which are divided into two main groups. On the one hand, there are those forests with a closed canopy at least 16.5 ft. (5 m) high. The canopy is the upper portion of the trees in the forest, and closed-canopy forests are so dense with vegetation that from the ground the sky is not visible. On the other hand, the UNESCO system encompasses open woodlands with a shorter, more sparse, and unclosed canopy. The first group tends to be tropical and subtropical (located at or near the equator), while the second typically is located in temperate and subpolar forests—that is, in a region between the two tropical latitudes and the Arctic and Antarctic circles, respectively. In the next paragraphs, we examine a few varieties of forest as classified by UNESCO.

TROPICAL AND SUBTROPICAL FORESTS. Tropical rain forests are complex ecosystems with a wide array of species. The dominant tree type is an angiosperm (a type of plant that produces flowers during sexual reproduction), known colloquially as tropical hardwoods. The climate and weather are what one would expect to find in a place called a tropical rain forest, that is, rainy and warm. When the rain falls, it cools things down, but when the sun comes back out, it turns the world of the tropical rain forest into a humid, sauna-like environment.

Naturally, the creatures that have evolved in and adapted to a tropical rain forest environment are those capable of enduring high humidity, but they are tolerant of neither extremely cool conditions nor drought. Within those parameters, however, exists one of the most biodiverse ecosystems on Earth: the tropical rain forest is home to an astonishing array of animals, plants, insects, and microorganisms. Indeed, without the tropical rain forest, terrestrial (land-based) animal life on Earth would be noticeably reduced.

In the tropics, by definition the four seasons to which we are accustomed in temperate zones—winter, spring, summer, and fall—do not exist. In their place there is a rainy season and a dry season, but there is no set point in the year at which trees shed their leaves. In a tropical and subtropical evergreen forest conditions are much drier than in the rain forest, and individual trees or tree species may shed their leaves as a result of dry conditions. All trees and species do not do so at the same time, however, so the canopy remains rich in foliage year-round—hence the term evergreen. As with a rain forest, the evergreen forest possesses a vast diversity of species.

In contrast to the two tropical forest ecosystems just described, a mangrove forest is poor in species. In terms of topography and landform, these forests are found in low-lying, muddy regions near saltwater. Thus, the climate is likely to be humid, as in a rain forest, but only organisms that can tolerate flooding and high salt levels are able to survive there. Mangrove trees, a variety of angiosperm, are suited to this environment and to the soil, which is poor in oxygen.

TEMPERATE AND SUBARCTIC FORESTS. Among the temperate and subarctic forest types are temperate deciduous forests, containing trees that shed their leaves seasonally, and temperate and subarctic evergreen conifer forests, in which the trees produce cones bearing seeds. These are forest types familiar to most people in the continental United

States. The first variety is dominated by such varieties as oak, walnut, and hickory, while the second is populated by pine, spruce, or fir as well as other types, such as hemlock.

Less familiar to most Americans outside the West Coast are temperate winter-rain evergreen broadleaf forests. These forests are dominated by evergreen angiosperms and appear in regions that have both a pronounced wet season and a summer drought season. Such forests can be found in southern California, where an evergreen oak of the *Quercus* genus is predominant. Even less familiar to Americans is the temperate and subpolar evergreen rain forest, which is found in the Southern Hemisphere. Occurring in a wet, frost-free ocean environment, these forests are dominated by such evergreen angiosperms as the southern beech and southern pine.

Angiosperms vs. Gymnosperms

Several times we have referred to angiosperms, a name that encompasses not just certain types of tree but also all plants that produce flowers during sexual reproduction. The name, which comes from Latin roots meaning "vessel seed," is a reference to the fact that the plant keeps its seeds in a vessel whose name emphasizes these plants' sexual-type reproduction: an ovary.

Angiosperms are a beautiful example of how a particular group of organisms can adapt to specific ecosystems and do so in a way much more efficient than did their evolutionary forebear. Flowering plants evolved only about 130 million years ago, by which time Earth had long since been dominated by another variety of seed-producing plant, the gymnosperm, of which pines and firs are an example. Yet in a relatively short period of time, from the standpoint of the earth sciences, angiosperms have gone on to become the dominant plants in the world. Today, about 80% of all living plant species are flowering plants.

ANGIOSPERM VS. GYMNOSPERM SEEDS. How did they do this? They did it by developing a means to coexist more favorably than gymnosperms with the insect and animal life in their ecosystems. Gymnosperms produce their seeds on the surface of leaflike structures, making the seeds vulnerable to physical damage and drying as the wind whips the branches back and forth. Furthermore, insects and other animals view gymnosperm seeds as a source of nutrition.

In an angiosperm, by contrast, the seeds are tucked away safely inside the ovary. Furthermore, the evolution of the flower not only has added a great deal of beauty to the world but also has provided a highly successful mechanism for sexual reproduction. This sexual reproduction makes it possible to develop new genetic variations, as genetic material from two individuals of differing ancestry come together to produce new offspring.

GYMNOSPERM POLLINATION. Gymnosperms reproduce sexually as well, but they do so by a less efficient method. In both cases, the trees have to overcome a significant challenge: the fact that sexual reproduction normally requires at least one of the individual plants to be mobile. Gymnosperms package the male reproductive component in tiny pollen grains, which are released into the wind. Eventually, the grains are blown toward the female component of another individual plant of the same species.

This method succeeds well enough to sustain large and varied populations of gymnosperms but at a terrific cost, as is evident to anyone who lives in a region with a high pollen count in the spring. A yellow dust forms on everything. So much pollen accumulates on window sills, cars, mailboxes, and roofs that only a good rain (or a car wash) can take it away, and one tends to wonder what good all this pollen is doing for the trees.

The truth is that pollination is wasteful and inefficient. Like all natural mechanisms, it benefit the overall ecosystem, in this case, by making nutrient-rich pollen grains available to the soil. Packed with energy, pollen grains contain large quantities of nitrogen, making them a major boost to the ecosystem if not to the human environment. But it costs the gymnosperm a great deal, in terms of chemical and biological energy and material, to produce pollen grains, and the benefits are much more uncertain.

Pollen *might* make it to the right female component, and, in fact, it will, given the huge amounts of pollen produced. Yet the overall system is rather like trying to solve an economic problem by throwing a pile of dollar bills into the air and hoping that some of the money lands in the right place. For this reason, it is no surprise

KEY TERMS

ABUNDANCE: A measure of the degree to which an ecosystem possesses large numbers of particular species. An abundant ecosystem may or may not have a wide array of different species. Compare with *complexity.*

ANGIOSPERM: A type of plant that produces flowers during sexual reproduction.

BIOGEOCHEMICAL CYCLES: The changes that particular elements undergo as they pass back and forth through the various earth systems and particularly between living and nonliving matter. The elements involved in biogeochemical cycles are hydrogen, oxygen, carbon, nitrogen, phosphorus, and sulfur.

BIOTA: A combination of all flora and fauna (plant and animal life, respectively) in a region.

CANOPY: The upper portion of the trees in a forest. In a closed-canopy forest the canopy (which may be several hundred feet, or well over 50 meters, high) protects the soil and lower areas from sun and torrential rainfall.

CARNIVORE: A meat-eating organism.

COMPLEXITY: A measure of the degree to which an ecosystem possesses a wide array of species. These species may or may not appear in large numbers. Compare with *abundance.*

COMPOUND: A substance made up of atoms of more than one element chemically bonded to one another.

DECOMPOSERS: Organisms that obtain their energy from the chemical breakdown of dead organisms as well as from animal and plant waste products. The principal forms of decomposer are bacteria and fungi.

DECOMPOSITION REACTION: A chemical reaction in which a compound is broken down into simpler compounds or into its constituent elements. On Earth, this often is achieved through the help of detritivores and decomposers.

DETRITIVORES: Organisms that feed on waste matter, breaking down organic material into inorganic substances that then can become available to the biosphere in the form of nutrients for plants. Their function is similar to that of decomposers, but unlike decomposers—which tend to be bacteria or fungi—detritivores are relatively complex organisms, such as earthworms or maggots.

that angiosperms gradually are overtaking gymnosperms.

ANGIOSPERM POLLINATION. The angiosperm overcomes its own lack of mobility by making use of mobile organisms. Whereas insects and animals pose a threat to gymnosperms, angiosperms actually put bees, butterflies, hummingbirds, and other flower-seeking creatures to work aiding their reproductive process. By evolving bright colors, scents, and nectar, the flowers of angiosperms attract animals, which travel from one flower to another, accidentally moving pollen as they do.

Because of this remarkably efficient system, animal-pollinated species of flowering plants do not need to produce as much pollen as gymnosperms. Instead, they can put their resources into other important functions, such as growth and greater seed production. In this way, the angiosperm solves its own problem of reproduc-

KEY TERMS CONTINUED

ECOSYSTEM: A community of interdependent organisms along with the inorganic components of their environment.

ELEMENT: A substance made up of only one kind of atom. Unlike compounds, elements cannot be broken chemically into other substances.

ENERGY TRANSFER: The flow of energy between organisms in a food web.

FOOD WEB: A term describing the interaction of plants, herbivores, carnivores, omnivores, decomposers, and detritivores in an ecosystem. Each consumes nutrients and passes it along to other organisms. Earth scientists typically prefer this name to *food chain,* an everyday term for a similar phenomenon. A food chain is a series of singular organisms in which each plant or animal depends on the organism that precedes or follows it. Food chains rarely exist in nature.

GEOSPHERE: The upper part of Earth's continental crust, or that portion of the solid earth on which human beings live and which provides them with most of their food and natural resources.

GYMNOSPERM: A type of plant that reproduces sexually through the use of seeds that are exposed, not hidden in an ovary, as with an angiosperm.

HERBIVORE: A plant-eating organism.

HYDROSPHERE: The entirety of Earth's water, excluding water vapor in the atmosphere but including all oceans, lakes, streams, groundwater, snow, and ice.

NICHE: A term referring to the role that a particular organism plays within its biological community.

OMNIVORE: An organism that eats both plants and other animals.

ORGANIC: At one time, chemists used the term *organic* only in reference to living things. Now the word is applied to most compounds containing carbon and hydrogen, thus excluding carbonates (which are minerals) and oxides, such as carbon dioxide.

PHOTOSYNTHESIS: The biological conversion of light energy (that is, electromagnetic energy) from the Sun to chemical energy in plants.

SYSTEM: Any set of interactions that can be set apart mentally from the rest of the universe for the purposes of study, observation, and measurement.

tion—and as a side benefit adds enormously to the world's beauty.

The Complexity of Ecosystems

The relationships between these two types of seed-producing plant and their environments illustrate, in a very basic way, the complex interactions between species in an ecosystem. Environmentalists often speak of a "delicate balance" in the natural world, and while there is some dispute as to how delicate that balance is—nature shows an amazing resilience in recovering from the worst kinds of damage—there is no question that a balance of some kind exists.

To put it another way, an ecosystem is an extraordinarily complex environment that brings together biological, geologic, hydrologic, and atmospheric components. Among these components are trees and other plants; animals, insects, and microorganisms; rocks, soil, minerals, and landforms; water in the ground and on the surface, flowing or in a reservoir; wind, sun, rain, moisture; and all the other specifics that make up weather and climate.

In the present context, we have not attempted to provide anything even approaching a comprehensive portrait of an ecosystem, drawing together all or most of the aspects described in the preceding paragraph. A full account of even the simplest ecosystem would fill an entire book. Given that level of complexity, it is safe to say that one should be very cautious before tampering with the particulars of an ecosystem. The essay on Ecology and Ecological Stress concerns what happens when such tampering occurs.

WHERE TO LEARN MORE

Beattie, Andrew J., and Paul Ehrlich. *Wild Solutions: How Biodiversity Is Money in the Bank.* New Haven, CT: Yale University Press, 2001.

Diamond, Jared M. *Guns, Germs, and Steel: The Fates of Human Societies.* New York: W. W. Norton, 1997.

The Ecosystems Center. Marine Biological Laboratory, Woods Hole, Massachusetts (Web site). <http://ecosystems.mbl.edu/>.

Ecotopia (Web site). <http://www.ecotopia.com/>.

Jordan, Richard N. *Trees and People: Forestland, Ecosystems, and Our Future.* Lanham, MD: Regnery Publishing, 1994.

Living Things: Habitats and Ecosystems (Web site). <http://www.fi.edu/tfi/units/life/habitat/habitat.html>.

Martin, Patricia A. *Woods and Forests.* Illus. Bob Italiano and Stephen Savage. New York: Franklin Watts, 2000.

Nebel, Bernard J., and Richard T. Wright. *Environmental Science: The Way the World Works.* Upper Saddle River, NJ: Prentice Hall, 2000.

Rybolt, Thomas R., and Robert C. Mebane. *Environmental Experiments About Life.* Hillside, NJ: Enslow Publishers, 1993.

The State of the Nation's Ecosystems (Web site). <http://www.us-ecosystems.org/>.

Sustainable Ecosystems Institute (Web site). <http://www.sei.org/>.

ECOLOGY AND ECOLOGICAL STRESS

CONCEPT

Ecology is the study of the relationships between organisms and their environments. As such, it is subsumed into the larger subject of ecosystems, which encompasses both living and nonliving components of the environment. As a study of the biological aspect of ecosystems, ecology is properly a part of the biological sciences rather than the earth sciences; however, in practice it is difficult to draw a line between the disciplines. This is especially the case inasmuch as the study of the environment involves such aspects as soil science, where earth sciences and ecology meet. This fact, combined with increasing concerns over ecological stresses, such as the increase of greenhouse gases in the atmosphere, warrants the consideration of ecology in an earth sciences framework.

HOW IT WORKS

Ecosystems, Biological Communities, and Ecology

An ecosystem is the complete community of living organisms and the nonliving materials of their surroundings. It therefore includes components that represent the atmosphere, the hydrosphere (all of Earth's waters, except for moisture in the atmosphere), the geosphere (the soil and extreme upper portion of the continental crust), and the biosphere. The biosphere includes all living things: plants (from algae and lichen to shrubs and trees); mammals, birds, reptiles, amphibians, aquatic life, and insects as well as all manner of microscopic forms, including bacteria and viruses. In addition, the biosphere draws together all formerly living things that have not yet decomposed.

The components of the biosphere are united not only by the fact that all of them are either living or recently living but also by the food web. The food web, discussed in much detail within the context of Ecosystems, is a complex network of feeding relationships and energy transfers between organisms. At various levels and stages of the food web are plants; herbivores, or plant-eating organisms; carnivores (meat-eating organisms); omnivores (organisms that eat both meat and plants); and, finally, decomposers and detritivores, which obtain their energy from the chemical breakdown of dead organisms.

ECOLOGY. When discussing the living components of an ecosystem—that is, those components drawn from the biosphere—the term biological community is used. This also may be called *biota,* which refers to all flora and fauna, or plant and animal life, respectively, in a particular region. The relationship between these living things and their larger environment, as we have noted, is called ecology. Pioneered by the German zoologist Ernst Haeckel (1834–1919), ecology was long held in disdain by the world scientific community, in part because it seemed to defy classification as a discipline. Though its roots clearly lie in biology, its broadly based, multidisciplinary approach seems more attuned to the earth sciences.

In any case, ecology long since has gained the respect it initially failed to receive, and much of that change has to do with a growing acceptance of two key concepts. On the one hand, there is the idea that all of life is interconnected and that the living world is tied to the nonliving, or inorganic, world. This is certainly a prevailing belief in the modern-day earth sciences, with its systems approach (see Earth Systems). On the other hand, there is the gathering awareness that certain aspects of industrial civilization may have a negative impact on the environment.

Clearly, the ecosystem as a whole is held together by tight bonds of interaction, but where the biological community is concerned, those bonds are even tighter. For the biological community to survive and thrive, a balance must be maintained between consumption and production of resources. Nature provides for that balance in numerous ways, but beginning in the late twentieth century, environmentalists in the industrialized world became increasingly concerned over the possibly negative effects their own societies exert on Earth's ecosystems and ecological communities.

Climax and Succession

One of the concerns raised by environmentalists is the issue of endangered species, or varieties of animal whose existence is threatened by human activities. In fact, nature itself sometimes replaces biological communities in a process called succession. Succession involves the progressive replacement of earlier biological communities with others over time. Coupled with succession is the idea of climax, a theoretical notion intended to describe a biological community that has reached a stable point as a result of ongoing succession.

Succession typically begins with a disturbance exerted on the preexisting ecosystem, and this disturbance usually is followed by recovery. This recovery may constitute the full extent of the succession process, at which point the community is said to have reached its climax point. Whether or not this happens depends on such particulars as climate, the composition of the soil, and the local biota.

There are two varieties of succession, primary and secondary. Primary succession occurs in communities that have never experienced significant modification of biological processes. In other words, the community affected by primary succession is "virgin," and primary succession typically involves enormous stresses. On the other hand, secondary succession happens after disturbances of relatively low intensity, such that the regenerative capacity of the local biota has not been altered significantly. Secondary succession takes place in situations where the biological community has experienced alteration.

Niche

Whereas climax and succession apply to broad biological communities, the term niche refers to the role a particular organism or species plays within the larger community. Though the concept of niche is abstract, it is unquestionable that each organism plays a vital role and that the totality of the ecosystem would suffer stress if a large enough group of organisms were removed from it. Furthermore, given the apparent interrelatedness of all components in a biological community, every species must have a niche—even human beings.

An interesting idea related to the niche is the concept of an indicator species: a plant or animal that by its presence, abundance, or chemical composition demonstrates a particular aspect of the character or quality of the environment. Indicator species can, for instance, be plants that accumulate large concentrations of metals in their tissues, thus indicating a preponderance of metals in the soil. This metal could indicate valuable deposits nearby, or it could serve as a sign that the soil is being contaminated.

In the rest of this essay, we explore a few examples of ecological stress—situations in which the relationship between organisms and environment has been placed under duress. We do not attempt to explore the ideas of succession, climax, niche, or indicator species with any consistency or depth; rather, our purpose in briefly discussing these terms is to illustrate a few of the natural mechanisms observed or hypothesized by ecologists in studying natural systems. The vocabulary of ecology, in fact, is as complex and varied as that of any natural science, and much of it is devoted to the ways in which nature responds to ecological stress.

REAL-LIFE APPLICATIONS

Deforestation

In Ecosystems, we discuss a number of forest types, whose makeup is determined by climate and the dominant tree varieties. Here let us consider what happens to a forest—particularly an old-growth forest—that experiences significant

RAIN FOREST DESTRUCTION BY FIRE IN MADAGASCAR. SUCH DEFORESTATION AFFECTS THE CARBON BALANCE IN THE ATMOSPHERE AND THE DIVERSITY OF SPECIES ON EARTH. *(© Daniel Heuclin/Photo Researchers. Reproduced by permission.)*

disturbance. Actually, the term deforestation can describe any interruption in the ordinary progression of the forest's life, including clear-cut harvesting, even if the forest fully recovers.

Deforestation can take place naturally, as a result of changes in the soil and climate, but the most significant cases of deforestation over the past few thousand years have been the result of human activities. Usually, deforestation is driven by the need to clear land or to harvest trees for fuel and, in some cases, building. Though deforestation has been a problem the world over, since the 1970s it has become more of an issue in developing countries.

DEVELOPED AND DEVELOPING NATIONS. In developed nations such as the United States, environmental activism has raised public awareness concerning deforestation and has led to curtailment of large-scale cutting in forests deemed important environmental habitats. By contrast, developing nations, such as Brazil, are cutting down their forests at an alarming rate. Generally, economics is the driving factor, with the need for new agricultural land or the desire to obtain wood and other materials driving the deforestation process.

Yet the deforestation of such valuable reserves as the Amazon rain forest is an environmental disaster in the making: as noted in Soil Conservation, the soil in rain forests is typically "old" and leached of nutrients. Without the constant reintroduction of organic material from the

OLD-GROWTH FORESTS ARE HOME TO THE NORTHERN SPOTTED OWL, RECOGNIZED AS AN ENDANGERED SPECIES BECAUSE OF THE DESTRUCTION OF ITS HABITAT. *(© T. Davis/Photo Researchers. Reproduced by permission.)*

plants and animals of the rain forests, it would be too poor to grow anything. Therefore, when nations cut down their own rain forest lands, in effect, they are killing the golden goose to get the egg. Once the rain forest is gone, the land itself is worthless.

CONSEQUENCES OF DEFORESTATION. Deforestation has several extremely serious consequences. From a biological standpoint, it greatly reduces biodiversity, or the range of species in the biota. In the case of tropical rain forests as well as old-growth forests, certain species cannot survive once the environmental structure has been ruptured. From an environmental perspective, it leads to dangerous changes in the carbon content of the atmosphere, discussed later in this essay. In the case of old-growth forests or rain forests, deforestation removes an irreplaceable environmental asset that contributes to the planet's biodiversity—and to its oxygen supply.

Even from a human standpoint, deforestation takes an enormous toll. Economically, it depletes valuable forest resources. Furthermore, deforestation in many developing countries often is accompanied by the displacement of indigenous peoples. Other political and social horrors sometimes lurk in the shadows: for example, Brazil's forests are home to charcoal plants that amount to virtual slave-labor camps. Indians are lured from cities with promises of high income and benefits, only to arrive and find that the situation is quite different from what was advertised. Having paid the potential employer for transportation to the work site, however, they are unable to afford a return ticket and must labor to repay the cost.

OLD-GROWTH FORESTS

Old-growth forests represent a climax ecosystem—one that has come to the end of its stages of succession. They are dominated by trees of advanced age (hence the name *old-growth*), and the physical structure of these ecosystems is extraordinarily complex. In some places the canopy, or "rooftop," of the forest is dense and layered, while in others it has gaps. Tree sizes vary enormously, and the forest is littered with the remains of dead trees.

An old-growth forest, by definition, takes a long time to develop. Not only must it have been free from human disturbance, but it also must have been spared various natural types of disturbance that bring about succession: catastrophic storms or wildfire, for instance. For this reason, most old-growth forests are rain forests in tropical and temperate environments. Among North American old-growth forests are those of the United States Pacific Northwest as well as those in adjoining regions of southwestern Canada.

THE SPOTTED OWL. These old-growth forests are home to a bird that, in the 1980s and 1990s, became well known both to environmentalists and to their critics: the northern spotted owl, or *Strix occidentalis caurina.* A nonmigratory bird, the spotted owl has a breeding pattern such that it requires large tracts of old-growth, moist-to-wet conifer forest—that is, a forest dominated by cone-producing trees—as its habitat. Given the potential economic value of old-growth forests in the region, the situation became one of heated controversy.

On the one hand, environmentalists insisted that the spotted owl's existence would be threatened by logging, and, on the other hand, representatives of the logging industry and the local community maintained that prevention of logging in the old-growth forests would cost jobs and livelihoods. The question was not an easy one, pitting the interests of the environment against those of ordinary human beings. By the early 1990s, the federal government had stepped in on the side of the environmentalists, having recognized the spotted owl as a threatened species under the terms of the U.S. Endangered Species Act of 1973. Nonetheless, controversy over the spotted owl—and over the proper role of environmental, economic, and political concerns in such situations—continues.

The Greenhouse Effect

Deforestation and other activities pose potential dangers to our atmosphere. In particular, such activities have led to an increasing release of greenhouse gases, which may cause the warming of the planet. As discussed in Energy and Earth, the greenhouse effect, in fact, is a natural process. Though it is typically associated, in the popular vocabulary at least, with the destructive impact of industrial civilization on the environment, it is an extremely effective mechanism whereby Earth makes use of energy from the Sun.

Rather than simply re-radiating solar radiation, Earth traps some of this heat in the atmosphere with the help of greenhouse gases, such as carbon dioxide. As in the case of most natural processes, however, if a little bit of carbon dioxide in the atmosphere is good, this does not mean that a lot is better.

As noted in the essay Carbon Cycle, all living things contain carbon in certain characteristic structures; hence, the term *organic* refers to this type of carbon content. Though carbon dioxide is not an organic compound, it is emitted by animals: they breathe in oxygen, which undergoes a chemical reaction in their carbon-based bodies, and, as a result, carbon dioxide is released. Plants, on the other hand, receive this carbon dioxide and, through a chemical process in their own cellular structures, take in the carbon while releasing the oxygen.

THE RESULT OF CUTTING MATURE FORESTS. Mature forests, such as those of the old-growth variety, contain vast amounts of carbon in the form of living and dead organic material: plants, animals, and material in the soil. Because this quantity is much greater than in a younger forest, when deforestation occurs in a mature forest ecosystem, the mature forest will be replaced by an ecosystem that contains much smaller amounts of carbon.

Ultimately, the carbon from the former ecosystem will be released to the atmosphere in the form of carbon dioxide. This will happen quickly, if the biomass of the forest is burned, or more slowly, if the timber from the forest is used for a long periods of time, for instance, in the building of houses or other structures.

Before humans began cutting down forests, Earth's combined vegetation stored some 990 billion tons (900 billion metric tons) of carbon, 90% of it appeared in forests. Today only about 616 billion tons (560 billion metric tons) of carbon are stored in Earth's vegetation, and the amount is growing smaller as time passes. At the same time, the amount of carbon dioxide in the atmosphere has increased from about 270 parts per million (ppm) in 1850 to about 360 ppm in 2000, and, again, the increase continues.

SHOULD WE BE WORRIED? Given this rise in atmospheric carbon dioxide as a result of deforestation—not to mention the more well-known cause, burning of fossil fuels—it is no wonder that atmospheric scientists and environmentalists are alarmed. Some of these scientists hypothesize that larger concentrations of carbon dioxide in the atmosphere will lead to increased intensity of the greenhouse effect. If this is true, it is possible that global warming will ensue, an eventuality that could have enormous implications for human survival. As a worst-case scenario, the polar ice cap (see Glaciology) could melt, submerging the cities of Earth.

Before succumbing to the sort of doomsday thinking and scaremongering for which many environmentalists are criticized, however, it is important to recognize that several contingencies are involved: *if* carbon dioxide in the atmosphere causes an increase in the intensity of the greenhouse effect, it *could* cause global warming. The fact is that despite a few mild winters at the end of the twentieth century, it is far from clear that the planet is warming. The winter of 1993, for instance, produced one of the worst blizzards that the eastern United States has ever seen.

As recently as the mid-1970s some environmentalists claimed that Earth actually is *cooling*—a response to a spate of cold winters in that period. The fact of the matter is that climate cycles are difficult to determine and require the

KEY TERMS

BIOACCUMULATION: The buildup of toxic chemical pollutants in the tissues of individual organisms.

BIOLOGICAL COMMUNITY: The living components of an ecosystem.

BIOMAGNIFICATION: The increase in bioaccumulated contamination at higher levels of the food web. Biomagnification results from the fact that larger organisms consume larger quantities of food—and, hence, in the case of polluted materials, more toxins.

BIOSPHERE: A combination of all living things on Earth—plants, mammals, birds, reptiles, amphibians, aquatic life, insects, viruses, single-cell organisms, and so on—as well as all formerly living things that have not yet decomposed.

BIOTA: A combination of all flora and fauna (plant and animal life, respectively) in a region.

CANOPY: The upper portion of the trees in a forest. In a closed-canopy forest, the canopy (which may be several hundred feet, or well over 50 meters, high) protects the soil and lower areas from sun and torrential rainfall.

CARNIVORE: A meat-eating organism.

CLIMAX: A theoretical notion intended to describe a biological community that has reached a stable point as a result of ongoing succession.

DECOMPOSERS: Organisms that obtain their energy from the chemical breakdown of dead organisms as well as from animal and plant waste products. The principal forms of decomposer are bacteria and fungi.

DECOMPOSITION REACTION: A chemical reaction in which a compound is broken down into simpler compounds or into its constituent elements. On Earth, this often is achieved through the help of detritivores and decomposers.

DETRITIVORES: Organisms that feed on waste matter, breaking organic material down into inorganic substances that then can become available to the biosphere in the form of nutrients for plants. Their function is similar to that of decomposers; however, unlike decomposers—which tend to be bacteria or fungi—detritivores are relatively complex organisms, such as earthworms or maggots.

ECOLOGY: The study of the relationships between organisms and their environments.

perspective of several centuries' worth of data (at least), rather than just a few years' worth. (See Glaciology for a discussion of the Little Ice Age, which took place just a few centuries ago.)

Nonetheless, it is important to be aware of the legitimate environmental concerns raised by the increased presence of carbon dioxide in the atmosphere due to human activities. Atmospheric scientists continue to monitor levels of greenhouse gases and to form hypotheses regarding the ultimate effect of such activities as deforestation and the burning of fossil fuels.

BIOACCUMULATION AND BIOMAGNIFICATION

As we have seen in a number of ways, one of the key concepts of ecological studies is also a core principle in the modern approach to the earth sciences. In both cases, there is the idea that a disturbance in one area can lead to serious conse-

KEY TERMS CONTINUED

ECOSYSTEM: A community of interdependent organisms along with the inorganic components of their environment.

ENERGY TRANSFER: The flow of energy between organisms in a food web.

FOOD WEB: A term describing the interaction of plants, herbivores, carnivores, omnivores, decomposers, and detritivores in an ecosystem. Each of these organisms consumes nutrients and passes them along to other organisms. Earth scientists typically prefer this name to food chain, an everyday term for a similar phenomenon. A food chain is a series of singular organisms in which each plant or animal depends on the organism that precedes or follows it. Food chains rarely exist in nature.

GEOSPHERE: The upper part of Earth's continental crust, or that portion of the solid earth on which human beings live and which provides them with most of their food and natural resources.

GREENHOUSE EFFECT: Warming of the lower atmosphere and surface of Earth. This occurs because of the absorption of long-wavelength radiation from the planet's surface by certain radiatively active gases, such as carbon dioxide and water vapor, in the atmosphere. These gases are heated and ultimately re-radiate energy to space at an even longer wavelength.

HERBIVORE: A plant-eating organism.

HYDROSPHERE: The entirety of Earth's water, excluding water vapor in the atmosphere, but including all oceans, lakes, streams, groundwater, snow, and ice.

NICHE: A term referring to the role that a particular organism plays within its biological community.

OMNIVORE: An organism that eats both plants and other animals.

ORGANIC: At one time chemists used the term organic only in reference to living things. Now the word is applied to most compounds containing carbon, with the exception of carbonates (which are minerals) and oxides, such as carbon dioxide.

SUCCESSION: The progressive replacement of earlier biological communities with others over time.

SYSTEM: Any set of interactions that can be set apart from the rest of the universe for the purposes of study, observation, and measurement.

quences elsewhere. The interconnectedness of components in the environment thus makes it impossible for any event or phenomenon to be truly isolated.

A good example of this is biomagnification. Biomagnification is the result of bioaccumulation, or the buildup of toxic chemical pollutants in the tissues of individual organisms. Part of what makes these toxins dangerous is the fact that the organism cannot process them easily either by metabolizing them (i.e., incorporating them into the metabolic system, as one does food or water) or by excreting them. Yet the organism ultimately does release some toxins—by passing them on to other members of the food web. This increase in contamination at higher levels of the food web is known as biomagnification.

THE PROCESS OF BIOMAGNIFICATION. Among the most prominent examples of chemical pollutants that are bioaccumulated are such pesticides as DDT (dichlorodiphenyl-trichloroethane). DDT is a chlorinated hydrocarbon (see Economic Geology) used as an insecticide. Because of its hydrocarbon base, DDT is highly soluble in oils—and in the fat of organisms. Once pesticides such as DDT have been sprayed, rain can wash them into creeks and, finally, lakes and other bodies of water, where they are absorbed by creatures that drink or swim in the water.

Atmospheric deposition, for instance, from industrial smokestacks or automobile emissions, is another source of toxins. Sludge from a sewage treatment plant can make its way into water sources, spreading all sorts of pollutants to the food web. Whatever the case, these toxins usually enter the food web by attaching to the smallest components. Particles of pollutant may stick to algae, which are so small that the toxin does little damage at this level of the food web. But even a small herbivore, such as a zooplankton, when it consumes the algae, takes in larger quantities of the pollutant, and thus begins the cycle of biomagnification.

By the time the toxin has passed from a zooplankton to a small fish, the amount of pollutant in a single organism might be 100 times what it was at the level of the algae. The reason, again, is that the fish can consume 10 zooplankton that each has consumed 10 algae. (These particular numbers, of course, are used simply for the sake of convenience.) By the time the toxins have passed on to a few more levels in the food web, they might be appearing in concentrations as great as 10,000 times their original amount.

DDT BIOMAGNIFICATION. For a period of about two decades before 1972, DDT was used widely in the United States to help control the populations of mosquitoes and other insects. Eventually, however, it found its way into water sources and fish species through the process we have described. Predatory birds, such as osprey, peregrine falcons, and brown pelicans, consumed these fish. So, too, did the bald eagle, which has long been a protected species owing to its role as America's national symbol.

DDT levels became so high that the birds' eggshells became abnormally thin, and adult birds sitting on nests accidentially would break the shells of unhatched eggs. As a result, baby birds died, and populations of these species also died. Public awareness of this phenomenon, raised by environmentalists in the late 1960s and early 1970s, led to the banning of DDT spraying in 1972. Since that time, populations of many predatory birds have increased dramatically.

ADDRESSING ECOLOGICAL CONCERNS. In the case of DDT biomagnification, humans were not directly involved, because the species of birds affected were not ones that people consume for food. Yet bioaccumulation and biomagnification have threatened humans. For example, in the 1950s, cows fed on grass that had been exposed to nuclear radiation and this radioactive material found its way into milk. Another example occurred during the 1970s and 1980s, when fish, such as tuna, were found to contain abnormally high levels of mercury.

This led the federal government and some states to issue warnings against the consumption of certain types of fish, owing to bioaccumulated levels of toxic pollutants. Obviously, such measures, however well intentioned, are just cosmetic fixes for larger problems. In the long run, what is needed is a systemic ecological approach that attempts to address problems such as biomagnification and the accumulation of greenhouse gases by approaching the root causes.

WHERE TO LEARN MORE

Ashworth, William, and Charles E. Little. *Encyclopedia of Environmental Studies.* New York: Facts on File, 2001.

The Ecological Society of America: Issues in Ecology (Web site). <http://esa.sdsc.edu/issues.htm>.

Ecology.com—An Ecological Source of Information (Web site). <http://www.ecology.com/>.

Ecology WWW Page (Web site). <http://pbil.univ-lyon1.fr/Ecology/Ecology-WWW.html>.

Endangered Species on EE-Link (Web site). <http://216.144.194.35/EndSpp/endangeredspecies-mainpage.html>.

Markley, O. W., and Walter R. McCuan. *21st Century Earth: Opposing Viewpoints.* San Diego, CA: Greenhaven Press, 1996.

The Marshall Cavendish Illustrated Encyclopedia of Plants and Earth Sciences. New York: Marshall Cavendish, 1988.

The Need to Know Library—Ecology and Environment Page (Web site). <http://www.peak.org/~mageet/tkm/ecolenv.htm>.

Philander, S. George. *Is the Temperature Rising?: The Uncertain Science of Global Warming.* Princeton, NJ: Princeton University Press, 1998.

Schultz, Warren. *The Organic Suburbanite: An Environmentally Friendly Way to Live the American Dream.* Emmaus, PA: Rodale, 2001.

WATER AND THE EARTH

HYDROLOGY

THE HYDROLOGIC CYCLE

GLACIOLOGY

HYDROLOGY

CONCEPT

Hydrology is among the principal disciplines within the larger framework of hydrologic sciences, itself a subcategory of earth sciences study. Of particular importance to hydrology is the hydrologic cycle by which water is circulated through various earth systems above and below ground. But the hydrologic cycle is only one example of the role water plays in the operation of earth systems. Outside its effect on living things, a central aspect of the hydrologic cycle, water is important for its physical and chemical properties. Its physical influence, exerted through such phenomena as currents and floods, can be astounding but no less amazing than its chemical properties, demonstrated in the fascinating realm of karst topography.

HOW IT WORKS

The Systems Approach

The modern study of earth sciences looks at the planet as a large, complex network of physical, chemical, and biological interactions. This is known as a systems approach to the study of Earth. The systems approach treats Earth as a combination of several subsystems, each of which can be viewed individually or in concert with the others. These subsystems are the geosphere, atmosphere, hydrosphere, and biosphere.

The geosphere is that part of the solid earth on which people live and from which are extracted the materials that make up our world: minerals and rocks as well as the organic products of the soil. In the latter area, the geosphere overlaps with the biosphere, the province of all living and recently living things; in fact, once a formerly living organism has decomposed and become part of the soil, it is no longer part of the biosphere and has become a component of the geosphere.

Overlap occurs between all spheres in one way or another. Thus, the hydrosphere includes all of the planet's waters, except for water that has entered the atmosphere in the form of evaporation. From the time moisture is introduced to the blanket of gases that surrounds the planet until it returns to the solid earth in the form of precipitation, water is a part of the atmosphere. This aspect of the planet's water is treated in the essay on Evapotranspiration and Precipitation.

THE HYDROSPHERE AND THE HYDROLOGIC CYCLE. All aspects of water on Earth, other than evaporation and precipitation, fall within the hydrosphere. This includes saltwater and freshwater, water on Earth's surface and below it, and all imaginable bodies of water, from mountain streams to underground waterways and from creeks to oceans. One of the fascinating things about water is that because it moves within the closed system of Earth, all the planet's water circulates endlessly. Thus, there is a chance that the water in which you take your next bath or shower also bathed Cleopatra or provided a drink to Charlemagne's horse.

On a less charming note, there is also a good chance that the water with which you brush your teeth once passed through a sewer system. Lest anyone panic, however, this has always been the case and always will be; as we have noted, water circulates endlessly, and one particular molecule may serve a million different functions.

Furthermore, as long as water continues to circulate through the various earth systems—that is, as long as it is not left to stagnate in a pond—it undergoes a natural cleansing process. Modern municipal and private water systems provide further treatment to ensure that the water that people use for washing is at least reasonably clean. In any case, it is clear that the movement of water through the hydrologic cycle is a subject complex enough to warrant study on its own (see Hydrologic Cycle).

The Hydrologic Sciences

As noted in Studying Earth, the earth sciences can be divided into three broad areas: the geologic, hydrologic, and atmospheric sciences. Each of these areas corresponds to one of the "spheres," or subsystems within the larger earth system, that we have discussed briefly: geosphere, hydrosphere, and atmosphere.

The hydrologic sciences are concerned with the hydrosphere and its principal component, water. These disciplines include glaciology—the study of ice in general and glaciers in particular—and oceanography. Glaciology is discussed in a separate essay, and oceanography is examined briefly in the present context. Aside from these two areas of study, the central component of the hydrologic sciences is hydrology—its most basic discipline, as geology is to the geologic sciences.

OCEANOGRAPHY. Oceanography is the study of the world's saltwater bodies—that is, its oceans and seas—from the standpoint of their physical, chemical, biological, and geologic properties. These four aspects of oceanographic study are reflected in the four basic subdisciplines into which oceanography is divided: physical oceanography, chemical oceanography, marine geology, and marine ecology. Each represents the application of a particular science to the study of the oceans.

Physical oceanography, as its name implies, involves the study of physics as applied to the world's saltwater bodies. In general, it concerns the physical properties of the oceans and seas, including currents and tides, waves, and the physical specifics of seawater itself—that is, its temperature, pressure at particular depths, density in specific areas, and so on.

Just as physical oceanography weds physics to the study of seawater, chemical oceanography is concerned with the properties of the ocean as viewed from the standpoint of chemistry. These properties include such specifics as the chemical composition of seawater as well as the role the ocean plays in the biogeochemical cycles whereby certain chemical elements circulate between the organic and inorganic realms (see Biogeochemical Cycles, Carbon Cycle, and Nitrogen Cycle).

The biogeochemical focus of chemical oceanography implies an overlap with geochemistry. Likewise, marine geology exists at the nexus of oceanography and geology, involving, as it does, such subjects as seafloor spreading (see Plate Tectonics), ocean topography, and the formation of ocean basins. Finally, there is the realm where oceanography overlaps with biology, a realm known as marine ecology or biological oceanography. This subdiscipline is concerned with the wide array of life-forms, both plant and animal, that live in the oceans as well as the food webs whereby they interact with one another.

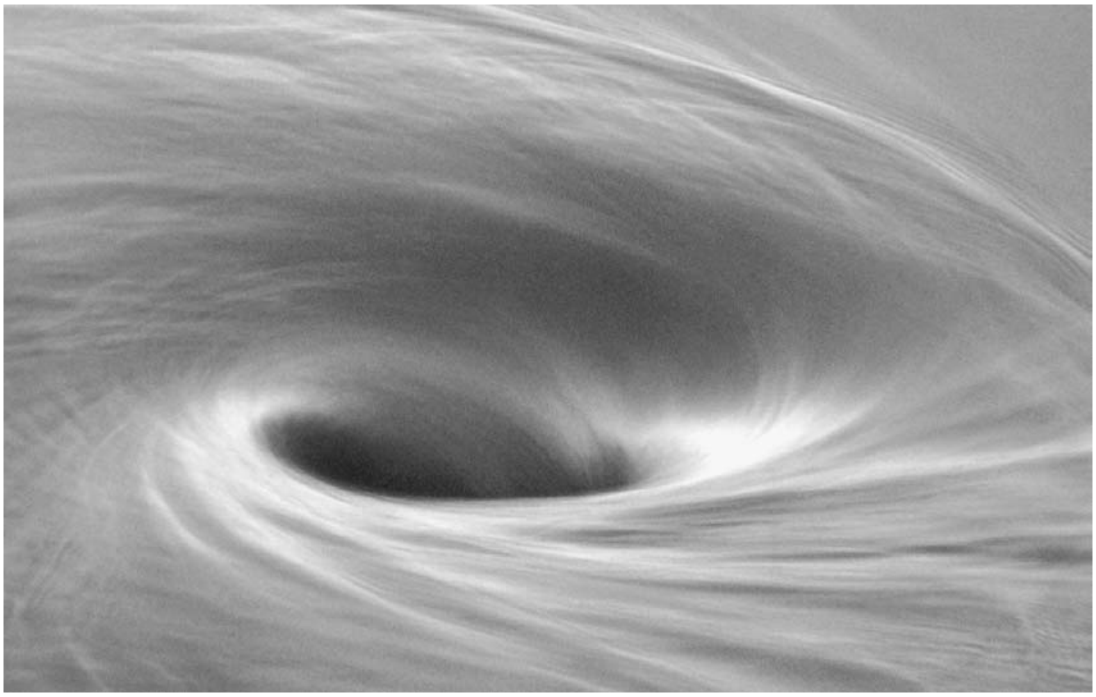
Introduction to Hydrology

As noted earlier, hydrology is the central field of the hydrologic sciences, dealing with the most basic aspects of Earth's waters. Among the areas of focus in hydrology are the distribution of water on the planet, its circulation through the hydrologic cycle, the physical and chemical properties of water, and the interaction between the hydrosphere and other earth systems.

Among the subdisciplines of hydrology are these:

- Groundwater hydrology: The study of water resources below ground.
- Hydrography: The study and mapping of large surface bodies of water, including oceans and lakes.
- Hydrometeorology: The study of water in the lower atmosphere, an area of overlap between the hydrologic and atmospheric sciences
- Hydrometry: The study of surface water—in particular, the measurement of its flow and volume.

THE WORK OF HYDROLOGISTS. Bringing together aspects of geology, chemistry, and soil science, hydrology is of enormous practical importance. Local governments, for instance, require hydrologic studies before the commencement of any significant building project, and hydrology is applied to such areas as the designation and management of flood plains. Hydrologists also are employed in the management of water resources, wastewater systems, and irrigation projects. The public use of water for



Whirlpools are created where two currents meet. Water tends to rotate in circles, clockwise in the Northern Hemisphere and counterclockwise in the Southern Hemisphere. *(© B. Tharp/Photo Researchers. Reproduced by permission.)*

recreation and power generation also calls upon the work of hydrologists, who assist governments and private companies in controlling and managing water supplies.

Hydrologists in the field use a variety of techniques, some of them simple and time-honored and others involving the most cutting-edge modern technology. They may make use of highly sophisticated computer models and satellite remote-sensing technology, or they may apply relatively uncomplicated methods for the measurement of snow depth or the flow of rivers and streams. Local hydrologists searching for water may even avail themselves of the services of quasi-mystics who employ a nonscientific practice called *dowsing.* The latter method, which involves the sensing of underground water with a "magic" divining rod, sometimes is used, with varying degrees of success, to find water in rural areas.

REAL-LIFE APPLICATIONS

Currents

Ocean waters are continually moving, not only as waves hitting the shore (a function of the Moon's gravitational pull—see Sun, Moon, and Earth) but also in the form of currents. These are patterns of oceanic flow, many of them regular and unchanging and others susceptible to change as a result of shifts in atmospheric patterns and other parameters. Among the factors that affect the flow of currents are landmasses, wind patterns, and the Coriolis effect, or the deflection of water caused by the turning of Earth.

Landmasses on either side of the Atlantic, Pacific, and Indian Oceans act as barriers to the paths of currents. For instance, if Africa were not placed as it is, between the Atlantic and Indian Oceans, water in the equatorial region would flow uniformly from east to west, or from the Indian Ocean to the Atlantic. Likewise, the movement would be uniformly west to east at the poles, as it would be at the equator.

Such is the case just off the shores of Antarctica, where the Antarctic Circumpolar Current, without the obstruction of land barriers, consistently circles the globe in a west-to-east direction. On the other hand, at the southern extremity of Africa, far from the equator, the movement of water also would be uniform, but in this case from west to east. Because of these landmasses, however, the movement of currents is much more complex.

Wind patterns also drive currents. These patterns work in tandem with the Coriolis effect, a term that generally describes a phenomenon that occurs with all particles on a rotating sphere such as Earth. The result of the Coriolis effect is that water tends to rotate in circles, clockwise in the Northern Hemisphere and counterclockwise in the Southern Hemisphere. (At the equator and poles, by contrast, the Coriolis effect is nonexistent.) Combined with prevailing winds, the Coriolis effect creates vast elliptical (oval-shaped) circulating currents called gyres.

SURFACE CURRENTS AND THEIR EFFECT ON CLIMATE. Among the basic types of currents are surface, tidal, and deep-water currents. In addition, a fourth type of current, a turbidity current, is of interest to oceanographers, hydrologists, and underwater geologists. Surface currents such as the Gulf Stream are the most well known variety, being the major form of current by which water circulates on the ocean's surface. Caused by the friction of atmospheric patterns—another type of current—moving over the sea, these currents largely are driven by winds. (Winds, in turn, are caused by differences in temperature between packets of air at differing altitudes—see Convection.)

Running as deep as 656 ft. (200 m), surface currents can be a powerful force. As a result of the Gulf Stream, for instance, a craft that sets sail eastward from the Caribbean is likely to be pulled quickly toward England. Of even greater importance is the impact of the Gulf Stream on climate. By moving warm waters in a northeasterly direction, it causes the European climate to be much warmer than it would be normally.

For instance, Boston, Massachusetts, and Rome, Italy, are on the same latitude, but whereas Boston is known for its icy winters, the mention of Rome rightly conjures images of sunshine and warmth. Likewise, London lies north of the 50th parallel, far above any city in the continental United States—including such places as International Falls, Minnesota, and Buffalo, New York, which are noted for their cruel winters. On the other hand, London, while it is far from balmy in wintertime, is many degrees warmer, thanks to the Gulf Stream.

OTHER TYPES OF CURRENT. Among the other types of currents are tidal currents, which are horizontal movements of water associated with the changing tides of the ocean. Their effect is felt primarily in the area between the continental shelf and the shore, where tidal-current phenomena such as riptides can pose a serious danger to swimmers. Deep-water currents, while they are less noticeable to people, are responsible for 90% of the water circulation that takes place in the ocean. Caused by variations in water density, which is a function of salt content (salinity) and temperature, these are slow-moving currents that move colder, denser water toward the depths of the ocean.

Then there are turbidity currents, which result from the mixing of relatively light water with water that has been made heavy by its sediment content. Earthquakes may cause these currents, which are local, fast movements of water along the ocean floor. Another cause of turbidity currents is the piling of sediment on underwater slopes. Turbidity currents play a major role in shaping the terrain of the ocean floor.

Flooding

Whereas currents arise in areas of Earth where water is "supposed" to be, floods, by definition, do not. They often occur in valleys or on coastlines and can be caused by various natural and man-made factors. Among natural causes are rains and the melting of snow and ice, while human-related causes can include poor engineering of irrigation or other water-management systems as well as the bursting of dams. In addition, the building of settlements too close to rivers and other bodies of water that are prone to flooding has resulted in the increase of human casualties from flooding over the centuries.

In terms of natural causes, changes in weather patterns typically are involved—but not always. For example, a low-lying coastal area may be susceptible to flooding at times when the ocean reaches high tide. (On the other hand, such weather conditions as low barometric pressure and high winds also can bring about heightened high tides.) Additionally, floods can be caused by earthquakes and other geologic phenomena that have no relation to the weather.

From ancient times people have located settlements near water. This settlement pattern resulted from the obvious benefits that accrued

THE FLOODWATERS OF THE MISSISSIPPI RISE TO 4 FT. (1.2 M), SURROUNDING THE PUMPING STATION IN HANNIBAL, MISSOURI. APART FROM NATURAL CAUSES, FLOODS CAN RESULT FROM INCONSISTENT FLOOD MANAGEMENT, POOR CIVIL ENGINEERING DESIGN, AND UNWISE AGRICULTURAL PRACTICES. *(AP/Wide World Photos. Reproduced by permission.)*

from access to water, and even though flooding was naturally a hazard, in some cases flooding itself was found to be beneficial. For the ancient Egyptians, the yearly cycles of flooding on the part of the Nile caused the deposition of rich soil, which played a major part in the fertility of the farmlands that, in turn, made possible the brilliant civilization of the pharaohs.

Along with these benefits, however, ancient peoples learned to fear the changes in weather and other circumstances that could bring about sudden flooding. This feeling is reflected, for instance, in Jesus' parable about the wise and foolish house builders. In the parable, a favorite Sunday school topic, the foolish man builds his house upon the sand, so that when the floods come, they sweep away his household. On the other hand, the wise man builds his house on rock, so that his household withstands the inevitable flood—an illustration about spiritual values that likewise reflects a reality of daily life in the ancient Near East.

HUMAN CAUSES AND EFFECTS. Humans can cause floods by such disastrous practices as clear-cutting of land and runaway grazing. Such activities remove vegetation, which holds soil in place and, in turn, keeps rivers and other bodies of water from flowing over onto the land. In addition, without vegetation to absorb rain, ground becomes saturated and thus susceptible to flooding. Not surprisingly, these unwise agricultural practices have helped bring about other disasters, such as the massive erosion of soil in the United States plains states that culminated in the dust bowl of the mid-1930s (see Soil Conservation).

Less well known than the dust bowl but still massive in its impact was the 1927 flood of the Mississippi River, which left more than a million people homeless. It, too, was in part the result of unwise practices, in this case, inconsistent flood management and civil engineering design, according to John M. Barry, the author of *Rising Tide: The Great Mississippi Flood of 1927 and How It Changed America.* As Barry indicated in his subtitle, the flood's impact went far beyond its direct effect on human lives or the landscape.

As Barry discusses in the book, the flood was a major cause behind the rise of the poor-white discontent in Louisiana that led to the governorship (and, in the opinion of some people, the dictatorship) of the notorious Huey P. Long (1893–1935). Long, who won election on promises to ease the suffering of the underclass, ultimately became the virtual ruler of his state, with a degree of power that in the opinions of some pundits rivaled that of President Franklin D. Roosevelt—if not that of his other contemporaries, Adolf Hitler and Benito Mussolini. Of even greater long-term significance, Barry maintained, the flood brought about the large-scale flight of African Americans to the north and the shift of black political allegiance from the Republican Party to the Democratic Party.

Few people alive today remember the 1927 flooding, but plenty recall the devastating floods of 1993, which killed 52 people and left over 70,000 homeless. Human mismanagement could not be blamed for the flooding itself, an outgrowth of exceptionally high soil moisture levels remaining from the fall of 1992, as well as heavy precipitation that continued in early 1993. However, once again, human attempts to control the flooding were less than successful: of some 1,300 levees or embankments that had been built (partly as a result of the 1927 flood) to keep flood waters back, all but about 200 failed. The floods, which lasted from late June to mid-August, destroyed nearly 50,000 homes and rendered over 12,000 sq mi. (31,000 sq km) of farmland useless. The overall damage estimate was in the range of $15 to $20 billion.

POWER AND PREVENTION. It is no wonder that a flood can have such a far-reaching impact, given the enormous power of water running wild in nature. Water is extremely heavy: just a bathtub full of water can weigh as much 750 lb. (340 kg). And it can travel as fast as 20 MPH (32 km/h), giving it tremendous physical force. Under certain conditions, a flood just 1 in. (2.54 cm) in depth can have as much potential energy as 60,000 tons (54,400 metric tons) of TNT. A U.S. study of persons killed in natural disasters during the 20-year period that ended in 1967 found that of 443,000 victims, nearly 40%, or about 173,000, were killed in floods. The other 60% was made up of people killed in 18 different types of other natural disaster, including hurricanes, earthquakes, and tornadoes.

Given this great destructive potential, communities—often with the help of hydrologists—have devised several means to control floods. Among such methods are the construction of dams and the diversion of floodwaters away from populated areas to flood-control reservoirs. These reservoirs then release the water at a slower rate than it would be released in the situation of a flood, thus giving the soil time to absorb the excess water. About one-third of all reservoirs in the United States are used for this purpose.

Hydrologists are particularly important in helping communities protect against flooding by methods known as hazard zoning and minimizing encroachment. By studying historical records, along with geologic maps and aerial photographs, hydrologists and other planners can make recommendations regarding the zoning laws for a particular area, so that builders will take special precautions. In addition, they can help minimize encroachment—that is, ensure that new buildings are not located in such a way that they restrict the flow of water or cause water to pool up excessively.

KARST TOPOGRAPHY

In contrast to the dramatic action of flooding or currents, karst topography is a more subtle but no less intriguing aspect of the ways in which water affects the dynamics of Earth. In this case, however, the effect is chemical rather than physical in origin. Karst topography is a particular variety of landscape created where water comes into contact with extremely soluble (easily dissolved in water) varieties of bedrock.

Karst is the German name for Kras, a region of Slovenia noted for its unusual landscape of strangely shaped white rock. In addition to the odd, funhouse forms of nightmarishly steep hills and twisting caves, much of it like something from a Dr. Seuss book, karst topography is noted for its absence of surface water, topsoil, or vegetation. The reason is that the bedrock comprises extremely soluble calcium carbonate minerals, such as limestone, gypsum, or dolomite.

Karst regions form as a result of chemical reactions between groundwater and bedrock. In the atmosphere and on the surface of the solid earth, water combines with carbon dioxide in the air, and this combination acts as a corrosive on calcium carbonate rocks. This corrosive or acidic material seeps into all crevices of the rock, developing into sinkholes and widening fissures over time. Gradually, it carves out enormous underground drainage systems and caves.

Sometimes the underground drainage structure collapses, leaving behind more odd shapes in the form of natural bridges and sink-

KEY TERMS

BIOGEOCHEMICAL CYCLES: The changes that particular elements undergo as they pass back and forth through the various earth systems and particularly between living and nonliving matter. The elements involved in biogeochemical cycles are hydrogen, oxygen, carbon, nitrogen, phosphorus, and sulfur.

BIOSPHERE: A combination of all living things on Earth—plants, mammals, birds, reptiles, amphibians, aquatic life, insects, viruses, single-cell organisms, and so on—as well as all formerly living things that have not yet decomposed.

CORIOLIS EFFECT: The deflection of water caused by the rotation of Earth. The Coriolis effect causes water currents to move in circles—clockwise in the Northern Hemisphere and counterclockwise in the Southern Hemisphere.

GEOCHEMISTRY: A branch of the earth sciences, combining aspects of geology and chemistry, that is concerned with the chemical properties and processes of Earth—in particular, the abundance and interaction of chemical elements and their isotopes.

GEOSPHERE: The upper part of Earth's continental crust, or that portion of the solid earth on which human beings live and which provides them with most of their food and natural resources.

HYDROLOGIC CYCLE: The continuous circulation of water throughout Earth and between various Earth systems.

HYDROLOGIC SCIENCES: Areas of the earth sciences concerned with the study of the hydrosphere. Among these disciplines are hydrology, glaciology, and oceanography.

HYDROLOGY: The study of the hydrosphere, including the distribution of water on Earth, its circulation through the hydrologic cycle, the physical and chemical properties of water, and the interaction between the hydrosphere and other earth systems.

HYDROSPHERE: The entirety of Earth's water, excluding water vapor in the atmosphere but including all oceans, lakes, streams, groundwater, snow, and ice.

ORGANIC: At one time, chemists used the term *organic* only in reference to living things. Now the word is applied to most compounds containing carbon, with the exception of carbonates (which are minerals) and oxides, such as carbon dioxide.

SYSTEM: Any set of interactions that can be set apart mentally from the rest of the universe for the purposes of study, observation, and measurement.

holes. This is a variety of karst topography known as doline karst. Another type is cone or tower karst, which produces tall, jagged limestone peaks, such as the sharp hills that characterize the river landscape in many parts of China. The United States is home to the world's largest karst region, which includes the Mammoth cave system in Kentucky.

WHERE TO LEARN MORE

Abbott, Patrick. *Natural Disasters.* Dubuque, IA: William C. Brown Publishers, 1996.

Arnell, Nigel. *Hydrology and Global Environmental Change.* Upper Saddle River, NJ: Prentice Hall, 2001.

Barry, John M. Rising Tide: *The Great Mississippi Flood of 1927 and How It Changed America.* New York: Simon and Schuster, 1997.

Dingman, S. Lawrence. *Physical Hydrology.* 2d ed. Upper Saddle River, NJ: Prentice Hall, 2002.

Gallant, Roy A. *Earth's Water.* New York: Benchmark Books, 2002.

Global Hydrology and Climate Center (Web site). <http://www.ghcc.msfc.nasa.gov/>.

Hydrology, Marine Science, Freshwater Science and Aquaculture Journals (Web site). <http://www.sciencekomm.at/journals/hydro.html>.

Karst Link Page (Web site). <http://hum.amu.edu.pl/~sgp/spec/linkk.html>.

Karst Waters Institute (Web site). <http://www.karstwaters.org/>.

Skinner, Brian J., Stephen C. Porter, and Daniel B. Botkin. *The Blue Planet: An Introduction to Earth System Science.* 2d ed. New York: John Wiley and Sons, 1999.

The World-Wide Web Virtual Library: Oceanography (Web site). <http://www.mth.uea.ac.uk/ocean/vl/>.

THE HYDROLOGIC CYCLE

CONCEPT

The hydrologic cycle is the continuous circulation of water throughout Earth and between Earth's systems. At various stages, water—which in most cases is synonymous with the hydrosphere—moves through the atmosphere, the biosphere, and the geosphere, in each case performing functions essential to the survival of the planet and its life-forms. Thus, over time, water evaporates from the oceans; then falls as precipitation; is absorbed by the land; and, after some period of time, makes its way back to the oceans to begin the cycle again. The total amount of water on Earth has not changed in many billions of years, though the distribution of water does. The water that we see, though vital to humans and other living things, makes up only about 0.0001% of the total volume of water on Earth; far more is underground and in other compartments of the environment.

HOW IT WORKS

Water and the Hydrosphere

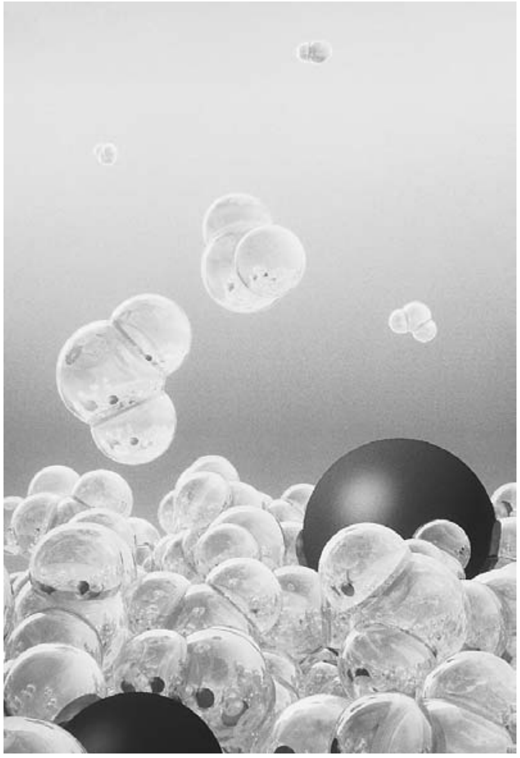
As we have noted, water and the hydrosphere are practically synonymous, but not completely so. The hydrosphere is the sum total of water on Earth, except for that portion in the atmosphere. This combines all water underground—which, as we shall see, constitutes the vast majority of water on the planet—as well as all freshwater in streams, rivers, and lakes; saltwater in seas and oceans; and frozen water in icebergs, glaciers, and other forms of ice (see Glaciology).

It is almost unnecessary to point out that water is essential to life. Human bodies, after all, are almost entirely made of water, and without water we would die much sooner than we would if we were denied food. Humans are not the only organisms dependent on water; whereas there are forms of life designated as *anaerobic,* meaning that they do not require oxygen, virtually nothing that lives can survive independent of water. Thus, the biosphere, which combines all living things and all recently deceased things, is connected intimately with the hydrosphere.

WATER ON EARTH. Throughout most of the modern era of scientific study—from the 1500s, which is to say most of the era of useful scientific study in all of human history—it has been assumed that water is unique to Earth. Presumably, if and when we found life on another planet, that planet also would contain water. But until that time, so it was assumed, we could be assured that the only planet with life was also the only planet with water.

In the latter part of the twentieth century, however, as evidence began to gather that Mars contains ice crystals on its surface, this exclusive association of water with Earth has been challenged. As it turns out, frozen water exists in several places within our solar system—as well it might, since water on Earth had to arrive from somewhere. It is believed, in fact, that water arrived on Earth at a very early stage, carried on meteors that showered the planet from space (see the entries Planetary Science and Sun, Moon, and Earth).

Since about three billion years ago, the amount of water on Earth has remained relatively constant. The majority of that water, however, is not in the biosphere, the atmosphere, or what we normally associate with the hydrosphere—



Water molecules evaporating from a solution. *(© K. Eward/Photo Researchers. Reproduced by permission.)*

that is, the visible rivers, lakes, oceans, and ice formations on Earth's surface. Rather, the largest portion of water on Earth is hidden away in the geosphere—that is, the upper portion of Earth's crust, on which humans live and from which we obtain the minerals and grow the plants that constitute much of the world we know.

Water Compartments in the Environment

In the course of circulating throughout Earth, water passes from the hydrosphere to the atmosphere. It does so through the processes of evaporation and transpiration. The first of these processes, of course, is the means whereby liquid water is converted into a gaseous state and transported to the atmosphere, while the second one—a less familiar term—is the process by which plants lose water through their stomata, small openings on the undersides of leaves. Earth scientists sometimes speak of the two as a single phenomenon, evapotranspiration.

Evaporation and transpiration, as well as the process whereby such moisture is returned to the solid earth—that is, precipitation—are discussed in the essay Evapotranspiration and Precipitation. Still, the atmosphere is just one of several "compartments" in which water is stored within the larger environment. Among the other important places in which water is found are the oceans and other surface waters, ice in its many forms, and aquifers. The latter are underground rock formations in which groundwater—water resources that occupy pores in bedrock—is stored.

IMBALANCES IN THE SYSTEM. The total amount of water in all these compartments is fixed, but water moves readily between various compartments through the processes of evaporation, precipitation, and surface and subsurface flows. The hydrologic cycle is thus a system all its own, a "system" (in scientific terms) being any set of interactions that can be set apart mentally from the rest of the universe for the purposes of study, observation, and measurement. Its net input and output balance each other. There may be imbalances of input and output in particular areas, which will manifest as drought or flooding.

Flooding, as well as other aspects of the hydrosphere and its study, is discussed in the essay Hydrology. As for drought, its immediate cause is a lack of precipitation, though other causes can be responsible for the removal of water from the local environment. For instance, at present a large portion of Earth's water is tied up in glaciers and other ice formations, but at other times in the planet's history this ice has been melted, leaving much of the continental mass that we know today submerged under water (see Glaciology).

ACCOUNTING FOR EARTH'S WATER. Earth's total water supplies are so large that instead of being measured by gallons or other units of volume, they are measured in terms of tons or metric tons, designated as *tonnes.* Nonetheless, for comparison's sake, consider the following figures in light of the fact that a gallon (3.8 l) of water weighs 8.4 lb. (3.8 kg). A ton contains 238 gal., and a tonne has 1,000 l.

Just as heat from the Sun accounts for the lion's share of Earth's total energy budget (see Energy and Earth), the vast majority of water on Earth comes from the deep lithosphere, the upper layer of Earth's interior, comprising the crust and the brittle portion at the top of the mantle. In this vast region are contained 2.76×10^{19} tons (2.5×10^{19} tonnes). This figure, equal to 27.6 billion billion tons, is about 94.7% of the global total.

The next largest compartment is the oceans, which contain 1.41×10^{18} tons (1.38×10^{18} tonnes), or 5.2% of the total. Ice caps, glaciers, and icebergs contain 1.74×10^{16} tons (0.017×10^{16} tonnes), thus accounting for most of the remaining 0.1% of Earth's water. Beyond these amounts are much smaller quantities representing shallow groundwater (2.76×10^{14} tons, or 2.5×10^{14} tonnes); inland surface waters, such as lakes and rivers (2.76×10^{13} tons, or 2.5×10^{13} tonnes); and the atmosphere (1.43×10^{13} tons, or 1.3×10^{13} tonnes).

REAL-LIFE APPLICATIONS

THE LIFE OF A WATER DROPLET

Now let us follow the progress of a single water droplet as it passes through the water cycle. This particular droplet, like all others, has passed through the cycle countless times over the course of the past few billion years, and in its various incarnations it has existed as groundwater, as moisture in the atmosphere, and as ice.

For the short span of Earth's existence that humans have occupied the planet, it is conceivable that our droplet has been consumed—either directly, as liquid water, or indirectly, as part of the water content in animal or vegetable material. That would mean that it also has been excreted, after which it will have continued the cycle of circulation. In theory, it might well be part of the water in which humans bathe, brush their teeth, or wash their clothes.

WATER AND FOREIGN MATERIAL. Of course, personal hygiene as we know it today is an extremely recent development: for instance, regular toothbrushing as a practice among the whole population began in the United States only around the turn of the nineteenth century. Still, it is a bit disconcerting to think that the water in which we brush our teeth today may have floated down a sewer pipe at another time. Nonetheless, by moving water through so many locales, the hydrologic cycle has a built-in cleansing component.

This cleansing component can be illustrated by the experience of saltwater, which despite its presence in the ocean is actually a small portion of Earth's total water supply. The reason is that the salt seldom travels with the water; as soon as the water evaporates, the salt is left behind. This is why people on the proverbial desert island or in other survival situations use evaporation to make saltwater drinkable.

Likewise, saltwater as such cannot survive the transition from liquid water to ice: as the water freezes, the ice (which has a much lower freezing point) simply is precipitated and left behind. Just as salt does not travel with water as it makes its way through the various stages of the hydrologic cycle, so other varieties of foreign matter are left behind as well; as long as water is not allowed to stagnate, it usually is cleansed in the course of traveling between the ground and the atmosphere.

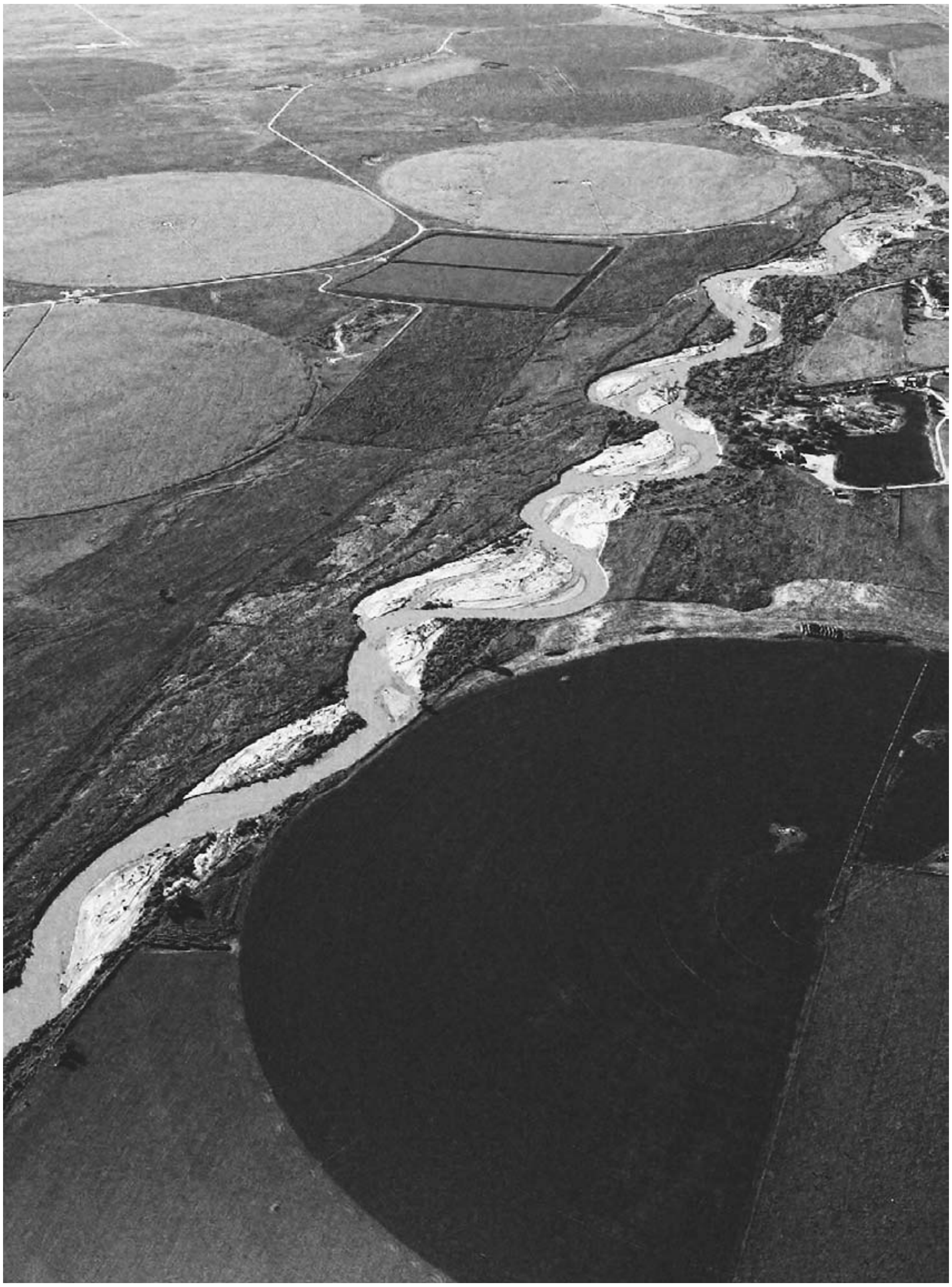
This is not to say that water typically exists in a pure form. Often called the *universal solvent,* water has such a capacity to absorb other substances that it is unlikely ever to appear in pure form unless it is distilled under laboratory conditions. Water in mountain streams, for instance, absorbs fragments of rock as it travels downhill, slowly eroding the surrounding rock and soil.

FROM THE WATERSHED TO POINTS BEYOND

On a particular watershed—an area of terrain from which water flows into a stream, river, or lake—our hypothetical water droplet may enter from a number of directions. In the simplest model of water flow, it comes from precipitation, including rain, snow, or even mist from clouds. The water has to go somewhere, and it may go either up or down. It may return to the atmosphere as evapotranspiration; it may enter the ground; or, if it reaches the solid earth at some elevation above sea level, it may enter a stream and flow ultimately to the ocean.

For water to enter the atmosphere generally requires an extensive surface area of vegetation, which supports high rates of transpiration. This transpiration, combined with evaporation from such inorganic surfaces as moist soil or bodies of water, puts a great deal of water into the atmosphere. Without significant evapotranspiration, however, it is necessary for the water to drain from the watershed, either as seepage to deep groundwater or as flow in the form of a stream.

THE FIVE STAGES OF THE HYDROLOGIC CYCLE. The overall process of the hydrologic cycle can be divided into five parts: condensation, precipitation, infil-



THE ARKANSAS RIVER, PART OF THE OGALLALA AQUIFER, RUNS THROUGH IRRIGATED FIELDS IN KANSAS. IRRIGATION PLACES A BIG DEMAND ON THE OGALLALA'S DIMINISHING WATER SUPPLY. *(AP/Wide World Photos. Reproduced by permission.)*

tration, runoff, and evaporation. Water vapor in the atmosphere condenses, forming clouds, which eventually become so saturated that they release the water to the solid earth in the form of precipitation. When precipitation enters the ground, it is known as infiltration.

Infiltration can be great or small, depending on the permeability of the ground. The soil of a rainforest, for instance, has so much organic matter that it is likely to be highly permeable. On the other hand, cities have large amounts of what land developers call impervious surface: roads, buildings, and other areas in which concrete and other materials prevent water from infiltrating the ground.

Assuming that water is unable to infiltrate, it becomes runoff. Runoff is simply surface water, which may take the form of streams, rivers, lakes, and oceans. If runoff occurs in an area that is not already a body of water, flood conditions may ensue. Thus, water may either infiltrate or become runoff, but as long as it remains close to the surface, it will experience evaporation.

In evaporation energy from the Sun changes liquid water into gaseous form, transporting it as a vapor into the atmosphere. Thus, the water is returned to the air, where it condenses and resumes the cycle we have described. As noted earlier, the water on or near Earth's surface is a small portion of the total. What about groundwater far below the surface? Let us now examine a particularly notable example of an aquifer, or groundwater reservoir.

The Ogallala Aquifer

Located beneath the central United States, the Ogallala Aquifer provides a vast store of groundwater that supports a large portion of American agriculture. The Ogallala, also known as the High Plains Regional Aquifer, was discovered in the early years of the nineteenth century. It did not become a truly significant economic resource, however, until the second half of the twentieth century, when advanced pumping technology made possible large-scale irrigation from the aquifer's supplies. By 1980 the Ogallala supported some 170,000 wells and accounted for fully one-third of all water used for irrigation purposes in the United States.

Centered in Nebraska, the aquifer underlies parts of seven other states: South Dakota, Wyoming, Colorado, Kansas, Oklahoma, Texas, and New Mexico. It stretches 800 mi. (1,287 km) from north to south and 200 mi. (322 km) from east to west at its widest point. All told, the Ogallala covers some 175,000 sq. mi. (453,250 sq km), an area larger than Germany—all of it underground.

In Nebraska, the aquifer is between 400 ft. and 1,200 ft. (130–400 m) deep, while at the southern edges its depth extends no more than 100 ft. (30 m). Composed of porous sand, silt, and clay formations deposited by wind and water from the Rocky Mountains, the Ogallala is made up of several sections, called formations. The largest of these is the Ogallala formation, which accounts for about 77% of its total volume.

USING UP THE OGALLALA. The Ogallala is particularly important to local agriculture because the states that it serves are home to numerous dry areas. Yet high-volume pumping of the underground reserves has reduced the available groundwater, much as the pumping of oil gradually is consuming Earth's fossil-fuel reserves. Indeed, the water of the Ogallala is known as fossil water, meaning that it has been stored underground for millions of years, just as the coal, oil, and gas that runs modern industrial civilization has.

Of course, the water from the Ogallala is not simply used up in an irreversible process, as is the case with fossil fuels; nonetheless, the rapidly accelerating reduction of its water supplies is cause for some alarm. Less than 0.5% of the water removed from this aquifer is being returned to the ground, and if the current rate of pumping increases, the supplies will be 80% depleted by 2020.

The consequences of this depletion are already being felt in Kansas, where streams and rivers, dependent on groundwater to feed their flow, are running dry. In that state alone, more than 700 mi. (1,126 km) of rivers that formerly flowed year-round have been reduced to dry channels. In New Mexico and Texas, use of center-pivot irrigation, which requires a well capable of pumping 750 gal. (2,839 l) per minute, is disappearing because the local aquifer can no longer sustain such volumes.

In addition to the problem of diminishing supplies, contamination is an issue. As more and more agricultural chemicals seep into an ever shrinking reservoir, the towns of the high plains—places once known for their pure, clean groundwater—now have tap water that is considered unsafe for children and pregnant women. Overuse of the Ogallala is also exacting a financial toll, as more and more wells run dry and farms go bankrupt.

KEY TERMS

AQUIFER: An underground rock formation in which groundwater is stored.

BEDROCK: The solid rock that lies below the C horizon, the deepest layer of soil.

BIOSPHERE: A combination of all living things on Earth—plants, mammals, birds, reptiles, amphibians, aquatic life, insects, viruses, single-cell organisms, and so on—as well as all formerly living things that have not yet decomposed.

EVAPORATION: The process whereby liquid water is converted into a gaseous state and transported to the atmosphere.

EVAPOTRANSPIRATION: The loss of water to the atmosphere via the processes of evaporation and transpiration.

GEOSPHERE: The upper part of Earth's continental crust, or that portion of the solid earth on which human beings live and which provides them with most of their food and natural resources.

GROUNDWATER: Underground water resources that occupy the pores in bedrock.

HYDROLOGIC CYCLE: The continuous circulation of water throughout Earth and between various earth systems.

HYDROLOGIC SCIENCES: Areas of the earth sciences concerned with the study of the hydrosphere. Among these areas of study are hydrology, glaciology, and oceanography.

HYDROLOGY: The study of the hydrosphere, including the distribution of water on Earth, its circulation through the hydrologic cycle, the physical and chemical properties of water, and the interaction between the hydrosphere and other earth systems.

HYDROSPHERE: The entirety of Earth's water, excluding water vapor in the atmosphere but including all oceans, lakes, streams, groundwater, snow, and ice.

LITHOSPHERE: The upper layer of Earth's interior, including the crust and the brittle portion at the top of the mantle.

PRECIPITATION: When discussing the hydrologic cycle or meteorology, precipitation refers to the water, in liquid or solid form, that falls to the ground when the atmosphere has become saturated with moisture.

SYSTEM: Any set of interactions that can be set apart mentally from the rest of the universe for the purposes of study, observation, and measurement.

TRANSPIRATION: The process whereby plants lose water through their stomata, small openings on the undersides of leaves.

WATERSHED: An area of terrain from which water flows into a stream, river, lake, or other large body.

Rivers

Despite the environmental challenges posed by such situations as the exhaustion of the Ogallala, the hydrologic cycle continues to roll on. As it does, it is sustained in large part by processes we cannot see: the movement of groundwater from the aquifer into streams or the evapotranspiration of surface waters to the atmosphere. Yet the movement of waters along the surface, because it is visible and recognizable to humans, attracts human attention in a way that many of these other components of the hydrologic cycle do not.

Rivers and other forms of surface water actually account for a relatively small portion of the planet's water supply, but they loom large in the human imagination as the result of their impact on our lives. The first human civilizations developed along rivers in Egypt, Mesopotamia, India, and China, and today many a great city lies along a river. Rivers provide us with a means of transportation and recreation, with hydroelectric power, and even—after the river water has been treated—with water for drinking and bathing.

FORMATION OF RIVERS. Rivers usually form from tributaries, such as springs. As the river flows, it is fed by more tributaries and by groundwater and continues on its way at various speeds, depending on the terrain. Finally, the river discharges into an ocean, a lake, or a desert basin.

River waters typically begin with precipitation, whether in the form of rainwater or melting snow. They also are fed by groundwater exuding from bedrock to the surface. When precipitation falls on ground that is either steeply sloped or already saturated, the runoff moves along Earth's surface, initially in an even, paper-thin sheet. As it goes along, however, it begins to form parallel rills, and its flow becomes turbulent. As the rills pass over fine soil or silt, they dig shallow channels, or runnels.

At some point in their flow, the runnels merge with one another, until there are enough of them to form a stream. Once enough streams have converged to create a continuously flowing body of water, it becomes a brook, and once the volume of water carried reaches a certain level, the brook becomes a river. As we have already noted, however, a river is really the sum of its tributaries, and thus hydrologists speak of river *systems* rather than single rivers.

RIVER SYSTEMS. A particularly impressive example of a river system is the vast Mississippi-Missouri, which drains the central United States. Most of the rivers between the Rockies and the Appalachians that do not empty directly into the Gulf of Mexico feed this system. This system includes the Ohio, itself an impressive river that divides the eastern United States. Indeed, just as the Mississippi separates east from west in America, the Ohio separates north from south.

After the Ohio and the Mississippi converge, at the spot where Illinois, Missouri, and Kentucky meet, they retain their separate identities for many miles. A strip of clear water runs along the river's eastern side, while to the west of this strip the water is a cloudy yellow—indicating a heavier amount of sediment in the Mississippi than in the Ohio. A similar phenomenon occurs where the "Blue Nile" and "White Nile," tributaries of another great river—both named for the appearance of the water—meet at Khartoum in Sudan.

WHERE TO LEARN MORE

"Groundwater Primer: Hydrologic Cycle." U.S. Environmental Protection Agency (Web site). <http://www.epa.gov/seahome/groundwater/src/cycle.htm>.

Hamilton, Kersten. *This Is the Ocean.* Honesdale, PA: Caroline House/Boyds Mills Press, 2001.

"Hydrologic Cycle." The Groundwater Foundation (Web site). <http://www.groundwater.org/GWBasics/hydro.htm>.

The Hydrologic Cycle (Web site). <http://agen521.www.ecn.purdue.edu/AGEN521/icon.html>.

Hydrologic/Water Cycle (Web site). <http://www.und.nodak.edu/instruct/eng/fkarner/pages/cycle.htm>.

Pidwirny, Michael J. "Introduction to Hydrology: The Hydrologic Cycle." *Fundamentals of Physical Geography* (Web site). <http://www.geog.ouc.bc.ca/physgeog/contents/8b.html>.

Richardson, Joy. *The Water Cycle.* Illus. Linda Costello. New York: F. Watts, 1992.

Smith, David. *The Water Cycle.* Illus. John Yates. New York: Thomson Learning, 1993.

Schaefer, Lola M. *This Is the Ocean.* Illus. Jane Wattenberg. New York: Greenwillow Books, 2001.

Walker, Sally M. *Water Up, Water Down: The Hydrologic Cycle.* Minneapolis: Carolrhoda Books, 1992.

GLACIOLOGY

CONCEPT

Glaciology is the study of ice and its effects. Since ice can appear on or in the earth as well as in its seas and other bodies of water and even its atmosphere, the purview of glaciologists is potentially very large. For the most part, however, glaciologists' attention is direction toward great moving masses of ice called glaciers, and the intervals of geologic history when glaciers and related ice masses covered relatively large areas on Earth. These intervals are known as ice ages, the most recent of which ended on the eve of human civilization's beginnings, just 11,000 years ago. The last ice age may not even be over, to judge from the presence of large ice masses on Earth, including the vast ice sheet that covers Antarctica. On the other hand, evidence gathered from the late twentieth century onward indicates the possibility of global warming brought about by human activity.

HOW IT WORKS

Ice

Ice, of course, is simply frozen water, and though it might appear to be a simple subject, it is not. Glaciologists classify differing types of ice, for instance, with regard to their levels of density, designating them with Roman numerals. The ice to which most of us are accustomed is classified as ice I. We will not be concerned with the other varieties of ice in the present context, but it should be noted that the ice in glaciers is quite different from the ice in an ice cube or even the ice on a pond in winter. These differences are a result of massive pressure, which reduces the air content of the ice in glaciers.

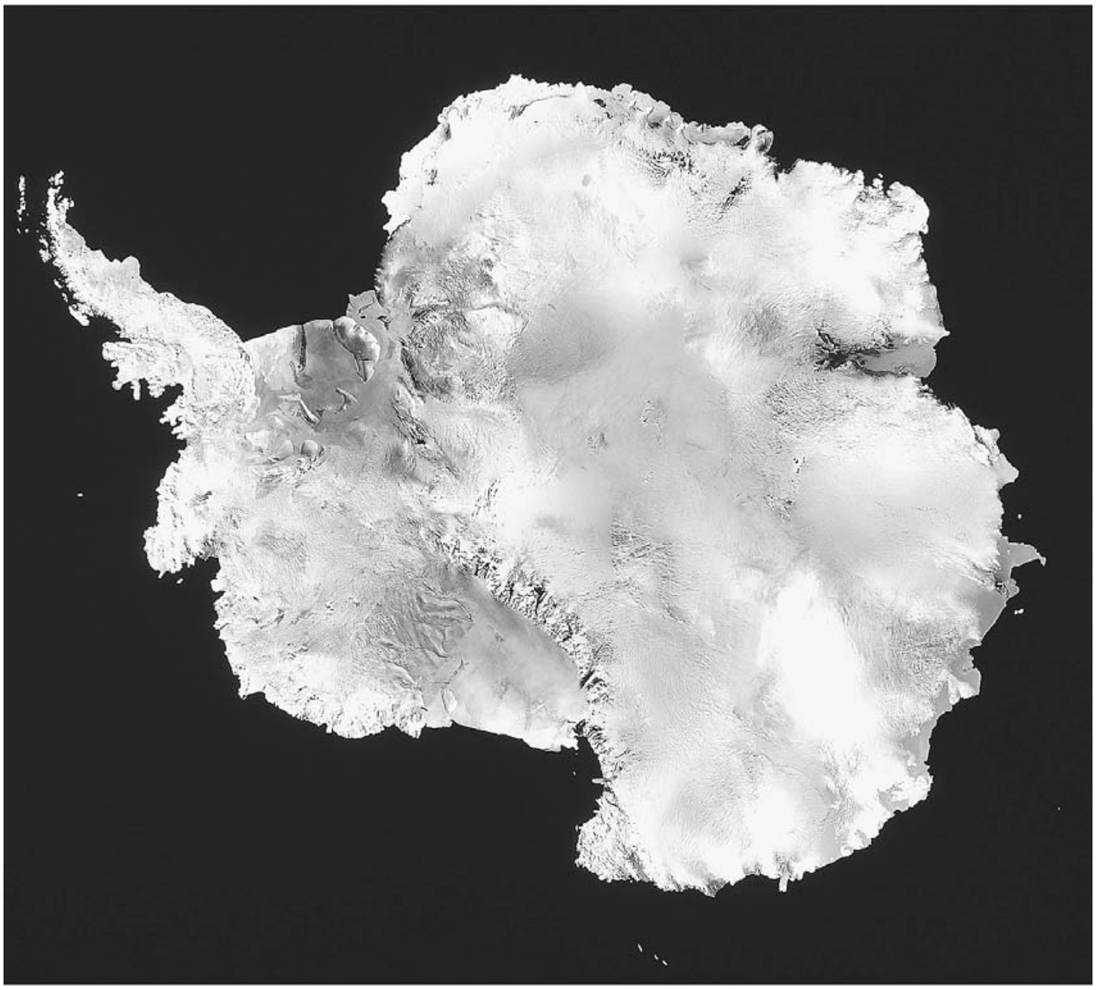
By definition, ice is composed of fresh water rather than saltwater. This is true even of icebergs, though they may float on the salty oceans. The reason is that water has a much higher freezing point than salt, and, therefore, when water freezes, very little of the salt remains joined to the water. Most of the salt is left behind in the form of a briny slush, and so much of Earth's fresh water supply actually is contained in great masses of ice, such as the glaciers of Antarctica.

Glaciology is defined as the study of ice, its forms, and its effects. This means that the glaciologist has a much wider scope than a geologist, meteorologist, or oceanographer, each of whom is concerned primarily with the geosphere, atmosphere, and hydrosphere, respectively. Though ice commonly is associated with the hydrosphere, where it appears on Earth's oceans, rivers, and lakes, it also is found on and even under the solid earth. There are even situations in which ice is found in the atmosphere.

Glaciology and Glaciers

Despite the wide distribution of ice on Earth and the many forms it takes, the work of most glaciologists is concerned primarily with ice as it appears in glaciers. A glacier is a large, typically moving mass of ice on or adjacent to a land surface. It does not flow, as water does; rather, it is moved by gravity, a consequence of its extraordinary weight.

Obviously, a glacier can form only in an extremely cold region—one so cold that the temperature never becomes warm enough for snow to melt completely. Some snow may melt as a result of contact with the ground, which is likely to be warmer than the snow itself, but when tem-



An infrared satellite image of Antarctica, showing ice shelves projecting from the coastline. The largest ice shelf, the Ross, sits to the left of the Transantarctic Mountains (lower center). *(© USGS/Photo Researchers. Reproduced by permission.)*

peratures drop, it refreezes. A glacier starts with a layer of ice, on which snow gathers until refreezing gradually creates compacted layers of snow and ice.

As anyone who has ever held a snowball in his or her hand knows, snow is fluffy, or, to put it in more scientific terms, it is much less dense than ice. A sample of snow is about 80% air, but as ice accumulates over a layer of snow, the weight of the ice squeezes out most of the air. As the layers grow thicker and thicker, the weight reduces the air further, creating an extremely dense, thick layer of ice. Ultimately, the ice becomes so heavy that its weight begins to pull it downhill, at which point it becomes a glacier.

Glacial Temperature and Morphologic Characteristics

Glaciers may be classified according to either relative temperature or morphologic characteristics (i.e., in terms of its shape). In terms of temperature, a glacier may be "warm," meaning that it is close to the pressure melting point. *Pressure melting point* is defined as the temperature at which ice begins to melt under a given amount of pressure. It is commonly known that water melts at 32°F (0°C), but only under conditions of ordinary atmospheric pressure at sea level. At higher pressures, the melting point of water is lower, which means that it can remain liquid at temperatures below its ordinary freezing point. (The melting point and freezing point of a substance are always the same.)

A "warm" glacier, such as those that appear in the Alps, is relatively mobile, because it is at the pressure melting point. This kind of glacier contrasts with a "cold," or polar, glacier, in which the temperatures are well below the pressure melting point; in other words, despite the extremely high pressure, the temperature is so low that the ice will not melt. As their name suggests, polar glaciers are found at Earth's poles, which effectively means Antarctica, since the area of the North Pole is not a land surface. A third category of glacier, in terms of temperature, is a subpolar glacier, found (not surprisingly) in regions near the poles. Examples of subpolar glaciers, or ones in which the fringes of the glacier are colder than the interior, are found in Spitsbergen, islands belonging to Norway that sit in the Arctic Ocean, well to the north of Scandinavia.

MORPHOLOGIC CLASSIFICATIONS. In the classification of geologic sciences, glaciology often is grouped with geomorphology. The latter field of study is devoted to landforms, or notable topographical features, and the forces and processes that have shaped them. Among those forces and processes are glaciers, which can be viewed in terms of their shape, the locale in which they form, and their effect on the contour of the land.

Alpine or mountain glaciers flow down a valley from a high mountainous region, typically following a path carved out by rivers or melting snow in warmer periods. They move toward valleys or the ocean, and in the process they exert considerable impact on the surrounding mountains, increasing the sharpness and steepness of these landforms. The rugged terrain in the vicinity of the Himalayas and the Andes, as well as the alpine regions of the Cascade Range and Rocky Mountains in the United States, are partly the result of weathering caused by these glaciers.

The glacial forms found in Alaska, Greenland, Iceland, and Antarctica are often piedmont glaciers, large mounds of ice that slope gently. Iceland, Greenland, and Antarctica as well as Norway are also home to cirque glaciers, which are relatively small and wide in proportion to their length. Though they experience considerable movement in place, they usually do not move out of the basinlike areas in which they are formed.

Other Ice Formations

There are several other significant varieties of ice formation, including ice caps, ice fields, and ice sheets. An ice cap, though much bigger than a glacier, typically has an area of less than 19,300 sq. mi. (50,000 sq km). Nonetheless, its mass is such that it exerts enormous weight on the land surface, and this exertion of force allows it to flow.

At the center of an ice cap or an ice sheet is an ice dome, and at the edges are ice shelves and outlet glaciers. Symmetrical and convex (i.e., like the outside of a bowl), an ice dome is a mass of ice often thicker than 9,800 ft. (3,000 m). An outlet glacier is a rapidly moving stream of ice that extends from an ice dome. Ice shelves, at the far outer edges, extend into the oceans, typically ending in cliffs as high as 98 ft. (30 m). Ice fields are similar to ice caps; the main difference is that the ice field is nearly level and lacks an ice dome. There are enormous variations in size for ice fields. Some may be no larger than 1.9 sq. mi. (5 sq km), while at different times in Earth's history, others have been as large as continents.

The most physically impressive of all ice formations, an ice sheet is a vast expanse of ice that gradually moves outward from its center. Ice sheets are usually at least 19,300 sq. mi. (50,000 sq km) and, like ice caps, consist of ice domes and outlet glaciers, with outlying ice shelves. Given their even greater size compared with ice caps, ice sheets exert still more force on the solid earth beneath them. They cause the rock underneath to compress, and, therefore, if an ice sheet ever melts, Earth's crust actually will rise upward in that area.

REAL-LIFE APPLICATIONS

Antarctica

An example of an ice sheet is the Antarctic ice sheet, which is permanently frozen—at least for the foreseeable future. The Antarctic ice sheet covers most of Antarctica, an area of about five million sq. mi. (12.9 million sq km), the size of the United States, Mexico, and Central America combined. Within it lies 90% of the world's ice and more fresh water than in all the planet's rivers and lakes combined. By contrast, the impressive Greenland ice sheet, at 670,000 sq. mi. (1,735,000 sq km), is dwarfed, as are smaller ice sheets in Iceland, northern Canada, and Alaska.

The Antarctic ice sheet is the Sahara of ice masses, though, in fact, it is almost 50% larger than the Sahara desert and a good deal more inhospitable. Whereas the Sahara is scattered with towns and oases and has a steady population of isolated villagers, nomads, and merchants in caravans, *no one* lives on the Antarctic ice sheet except scientists on temporary missions. And whereas people have lived in the Sahara for thousands of years (it became a desert only somewhat recently, during the span of human civilization), scientific missions to Antarctica became possible only in the twentieth century. As it is, researchers spend only short periods of time on the continent and then in heavily protected environments.

ICE SHELVES AND GLACIERS. Just as the Antarctic ice sheet is the largest in the world, one of its attendant shelves also holds first place among ice shelves. The continent is shaped somewhat like a baby chick, with its head and beak pointing northward toward the Falkland Islands off the coast of South America and its two greatest ice shelves lying on either side of the "neck."

Facing the Weddell Sea, and the southern Atlantic beyond it, is the Ronne Ice Shelf, which extends about 400 mi. (640 km) over the water. The world record–holder, however, is the Ross Ice Shelf on the other side of the "neck," near Marie Byrd Land. About the same size as Texas or Spain, the Ross shelf extends some 500 mi. (800 km) into the sea and is the site of several permanent research stations.

ANTARCTIC TOPOGRAPHY. The Antarctic is also home to a vast mountain range, the Transantarctic, which stretches some 3,000 mi. (4,828 km) across the "neck" between the Ross and Weddell seas. Included in the Transantarctic Mountains is Vinson Massif, which at 16,860 ft. (5,140 m) is the highest peak on the continent. The continent as a whole is largely covered with mountain ranges, between which lie three great valleys called the Wright, Taylor, and Victoria valleys. Each is about 25 mi. (40 km) long and 3 mi. (5 km) wide.

These are the largest continuous areas of ice-free land on the continent, and they offer rare glimpses of the rocks that form the solid-earth surface deep beneath Antarctica. They are also among the strangest places on the planet, forbidding lands even by Antarctic standards. The three are known as the "dry valleys," owing to their lack of precipitation; indeed, if they lay in a more temperate zone, they would be deserts far more punishing than the Sahara. Geologists estimate that it has not rained or snowed in these three valleys for at least one million years. The reason is that ceaseless winds keep the air so dry that any falling snow evaporates before it reaches the ground. In this arid, brutally cold climate, nothing decomposes, and seal carcasses a millennium old remain fully intact.

THE THICKNESS OF THE ICE. The dry valleys are exceptional, because most of Antarctica lies under so much ice that the rocks cannot be seen. The ice in Antarctica has an *average* depth of more than a mile: the depth averages about 6,600 ft. (2,000 m), but in places on the continent it is as thick as 2 mi. (3.2 km). Thus, "ground level" on Antarctica is equivalent to a fairly high elevation in the inhabitable portion of the planet. Denver, Colorado, for instance, touts itself as the "Mile-High City," and its elevation has enough effect on a visiting flatlander that rival sports teams usually spend a few days in Denver before a game, adjusting themselves to the altitude.

The thickness of the ice has allowed glaciologists to take deep ice-core samples from Antarctica. An ice core is simply a vertical section of ice that, when studied with the proper techniques and technology, can reveal past climatic conditions in much the same way that the investigation of tree rings does. (See Paleontology for more about tree-ring research, or dendrochronology.) Ice-core samples from the Antarctic provide evidence regarding Earth's climate for the past 160,000 years and show a pattern of warming and cooling that is related directly to the presence of carbon dioxide and methane in the atmosphere. These core samples also reveal the warming effects of increases in both gases over the past two centuries.

Because of the great thickness of its ice, Antarctica has the highest average elevation of any continent on the planet. Yet beneath all that ice, the actual landmass is typically well below sea level. The reason for this is that the ice weighs it down so much; by contrast, if the ice were to melt, the land would begin to spring upward. The melting of the Antarctic ice shelf would be a disaster of unparalleled proportions. If all that ice were to melt at once, it would raise global sea levels by some 200 ft. (65 m). This would be enough to flood all the world's ports, along with vast areas of low-lying land. For instance, waters would swell over New York City and all ports on America's eastern seaboard and probably would cover an area extending westward to the Appalachian Mountains. Even if only 10% of Antarctica's ice were to melt, the world's sea level would rise by 20 ft. (6 m), enough to cause considerable damage.

What Glaciers Do to Earth's Surface

Periodically over the past billion years, Earth's sea levels have advanced or retreated dramatically in conjunction with the beginning and end of ice ages. The latter will be discussed at the conclusion of this essay; in the present context, let us consider simply the geomorphologic effects of glaciers and ice masses. For example, as suggested in the discussion of Antarctica's ice sheet, when glaciers melt, thus redistributing their vast weight, Earth's crust rebounds. At the end of the last ice age, the crust rose upward, and in parts of North America and Europe this process of crustal rebounding is still occurring.

Glaciers move at the relatively slow speeds one would expect of massive objects made from ice: only a few feet or even a few inches per day. Friction with Earth's surface may melt the layer of ice that comes in contact with it, however, and, as a result, this layer of meltwater becomes like a lubricated surface, allowing the glacier to move much faster. The entire body of ice experiences a sudden increase of speed, called surging.

PLOWING THROUGH THE LAND. A glacier is like a huge bulldozer, plowing though rock, soil, and plants and altering every surface with which it comes into contact. It erodes the bottoms and sides of valleys, changing their V shape to a U shape. The rate at which it erodes the land is directly proportional to the depth of the glacier: the thicker the ice, the more it bears down on the land below it. As it moves, the ice pulls along rocks and soil, which are incorporated into the glacier itself. These components, in turn, make the glacier even more formidable, giving it greater weight, cutting ability, and erosive power.

The sediments left by glaciers that lack any intervening layer of melted ice are known by the general term till. In unglaciated areas, or places that have never experienced any glacial activity, sediment is formed by the weathering and decomposition of rock. On the other hand, formerly glaciated areas are distinguished by layers of till from 200 to 1,200 ft. (61–366 m) thick. Piles of till left behind by glaciers form hills called moraines, and the depressions left by these land-scouring ice masses are called kettle lakes.

North America abounds with examples of moraines and kettle lakes. Illinois, for instance, is covered with ridges, called end moraines, left behind by the melting near the conclusion of the last ice age. Visitors can take in a splendid view of moraine formations at Moraine View State Recreation Area, located astride the Bloomington moraine in central Illinois. Likewise Minnesota, Wisconsin, the Dakotas, and Wyoming are home to many a moraine. As with the Illinois recreation area, Kettle Lakes Provincial Park, near Timmins, Ontario, provides an opportunity to glimpse gorgeous natural wonders left behind by the retreat of glaciers—in this case, more than 20 deep kettle lakes. Park literature invites visitors to boat, fish, or swim in the lakes, though it would take a hardy soul indeed to brave those icy waters.

The glacier transports material from the solid earth as long as it is frozen, but wholly or partially melted glaciers leave behind sedimentary forms with their own specific names. In addition to moraines, there are piles of sediment, called eskers, left by rivers flowing under the ice. In addition, deposits of sediment may wash off the top of a glacier to form steep-sided hills called kames. If the glacier runs over moraines, eskers, or kames left by another glacier, the resulting formation is called a drumlin.

Just as rivers consist of main bodies formed by the flow of tributaries (for example, the many creeks and smaller rivers that pour into the Mississippi), so there are tributary glaciers. When a glacial tributary flows into a larger glacier, their top elevations become the same, but their bottoms do not. As a result, they carve out "hanging valleys," often the site of waterfalls. Examples include Yosemite and Bridalveil in California's Yosemite Valley.

The Kennicott glacier in the Wrangell Mountains of Alaska. *(© Pat and Tom Leeson/Photo Researchers. Reproduced by permission.)*

Ice Ages

The glaciers that exist today are simply the remnants of the last ice age, a time in which the size of the ice masses on Earth dwarfed even the great Antarctic ice sheet. When people speak of "the Ice Age," what they mean is the *last* ice age, which ended about 11,000 years ago. Yet it is one of only about 20 ice ages that have taken place over the past 2.5 million years, roughly coinciding with the late Pliocene and Pleistocene epochs. Actually, periods of massive glaciation (the covering of the landscape with large expanses of ice) have occurred at intervals over the past billion years. Their distribution over time has not been random; rather, they are concentrated at specific junctures in Earth's history.

Like the great mass extinctions of the past (see Paleontology), ice ages are among the markers geologists use in separating one interval of geologic time from another. In fact, there have been connections between ice ages and mass extinctions, particularly those that resulted from a recession of the seas. For example, the mass extinction that took place near the end of the Ordovician period (about 435 million years, or Ma, ago) came about as a result of a drop in the ocean level, which was caused, in turn, by an increase of glaciation that coincided with that phase in Earth's history.

The late Ordovician/early Silurian ice ages (between 460 to 430 Ma ago) are among four major phases of glaciation during the past 800

KEY TERMS

ATMOSPHERE: In general, an atmosphere is a blanket of gases surrounding a planet. Unless otherwise identified, however, the term refers to the atmosphere of Earth, which consists of nitrogen (78%), oxygen (21%), argon (0.93%), and other substances that include water vapor, carbon dioxide, ozone, and noble gases such as neon, which together comprise 0.07%.

CRUSTAL REBOUNDING: An upward movement by Earth's crust in response to the melting of a glacier, which redistributes its vast weight and causes Earth to rebound.

GEOMORPHOLOGY: An area of physical geology concerned with the study of landforms, with the forces and processes that have shaped them, and with the description and classification of various physical features on Earth.

GEOSPHERE: The upper part of Earth's continental crust, or that portion of the solid Earth on which human beings live and that provides them with most of their food and natural resources.

GLACIATION: The covering of the landscape with large expanses of ice, as during an ice age.

GLACIER: A large, typically moving mass of ice on or adjacent to a land surface.

GLACIOLOGY: An area of physical geology devoted to the study of ice, its forms, and its effects.

HYDROSPHERE: The entirety of Earth's water, excluding water vapor in the atmosphere but including all oceans, lakes, streams, groundwater, snow, and ice.

ICE AGE: A period of massive and widespread glaciation. Ice ages usually occur in series over stretches of several million years, or even several hundred million years.

ICE CAP: An ice formation bigger than a glacier but smaller than an ice sheet. An ice cap typically has an area of less than 19,300 sq. mi. (50,000 sq km) and, like an ice sheet, consists of an ice dome, with ice shelves and outlet glaciers at the edges.

ICE CORE: A vertical section of ice, usually taken from a deep ice sheet such as that in Antarctica. When studied with the proper techniques and technology, ice cores can reveal past climatic conditions in much the same way that the investigation of tree rings does.

ICE DOME: A symmetrical, convex (i.e., like the outside of a bowl) mass of ice, often thicker than 9,800 ft. (3,000 m), usually found at the center of an ice cap or an ice sheet.

million years. The first of these occurred during the late Proterozoic eon, toward the end of Precambrian time (between 800 and 600 Ma ago). Another happened during the Pennsylvanian subperiod of the Carboniferous and extended throughout the Permian period, thus lasting from about 350 to 250 million years ago. The last period of glaciation is the one in which we are living, beginning in the late Neogene period and extending into the current Quaternary.

HUMANS AND THE ICE AGES. In fact, many scientists question whether the last ice age has ended and whether we are merely living in an interglacial period. Certainly crustal rebounding is still taking place, as noted earlier.

KEY TERMS CONTINUED

ICE FIELD: A large ice formation, similar to an ice cap except that it is nearly level and lacks an ice dome. There are enormous variations in size for ice fields. Some may be no larger than 1.9 sq. mi. (5 sq km), while at different times in Earth's history, some have been as large as continents.

ICE SHEET: A vast expanse of ice, usually at least 19,300 sq. mi. (50,000 sq km), that moves outward from its center. Like the smaller ice caps, ice sheets consist of ice domes and outlet glaciers, with outlying ice shelves.

ICE SHELF: An ice formation at the edge of an ice cap or ice sheet that extends into the ocean, typically ending in cliffs as high as 98 ft. (30 m).

LANDFORM: A notable topographical feature, such as a mountain, plateau, or valley.

MA: An abbreviation used by earth scientists, meaning million years or megayears. When an event is designated as, for instance, 160 Ma, it usually means 160 million years ago.

MASS EXTINCTION: A phenomenon in which numerous species cease to exist at or around the same time, usually as the result of a natural calamity.

MORAINE: A hill-like pile of till left behind by a glacier.

MORPHOLOGY: Structure or form, or the study thereof.

OUTLET GLACIER: A rapidly moving stream of ice that extends from an ice dome.

PHYSICAL GEOLOGY: The study of the material components of Earth and of the forces that have shaped the planet. Physical geology is one of two principal branches of geology, the other being historical geology.

PRESSURE MELTING POINT: The temperature at which ice begins to melt under a given amount of pressure. The higher the pressure, the lower the temperature at which water can exist in liquid form.

RELIEF: Elevation and other inequalities on a land surface.

SEDIMENT: Material deposited at or near Earth's surface from a number of sources, most notably preexisting rock.

TILL: A general term for the sediments left by glaciers that lack any intervening layer of melted ice.

TOPOGRAPHY: The configuration of Earth's surface, including its relief as well as the position of physical features.

Inasmuch as "glacial period" refers to a time when glaciers cover significant portions of Earth and when the oceans are not at their maximum levels, we are indeed still in a glacial period.

It seems as though human existence has been bounded by ice, both in its onset and its recession. Ice ages have been a regular feature of the two million years since *Homo sapiens* came into existence, and the species had much of its formative experience in times of glaciation. The latter part of the last ice age created a land bridge that made possible the migration of Siberian peoples to the Americas, so that they are known now as *Native Americans.* (The name is well deserved: the ancestors of the Native Americans moved east from Siberia about 12,000 years ago, whereas less than half that much time has elapsed since the Indo-European ancestors of Caucasian Europeans moved west from what is now Russia. Certainly no one today questions whether Germans, Italians, British, French, and other groups are "native" Europeans.)

As an indication that ice ages have not ceased to occur, there is the Little Ice Age, which lasted from as early as 1250 to about 1850. This was a period of cooling and expansion of glaciers in the temperate latitudes on which Europe is located. Glaciers destroyed farmlands and buildings in the Alps, Norway, and Iceland, while Norse settlements in Greenland became uninhabitable. Europe as a whole suffered widespread crop failures, with a resulting loss of life. Evidence for this ice age, and indeed for all ice ages, can be discerned from the "footprints" left by glaciers from that time. Another telling sign is the transport of materials, such as rocks and fossils, from one part of Earth's surface to another.

UNDERSTANDING ICE AGES. What caused the Little Ice Age? The answer, or rather the attempt at an answer, goes to the core question of what causes ice ages in general. Earth scientists have cited both extraterrestrial and terrestrial factors. Among the leading extraterrestrial causes are an increase in sunspot activity (see Sun, Moon, and Earth for more about sunspots) as well as changes in Earth's orientation with respect to the Sun.

Contenders for a terrestrial explanation include changes in ocean circulation, as well as meteorites and volcanism. Either of these last two could have caused the atmosphere to become glutted with dust, choking out the Sun's light and cooling the planet considerably. (Such a calamity has been blamed for several instances of mass extinction, most notably, the one that wiped out the dinosaurs some 65 million years ago. See Paleontology.)

THE FUTURE. Though it appears that we are living in an interglacial period and that Earth could undergo significant cooling again thousands of years from now, there is also an even more frightening prospect of human-induced *warming.* As noted earlier, the Antarctic ice core reveals an increase of carbon dioxide and methane in the atmosphere during the past two centuries. Though these gases can be produced naturally, this excess in recent times appears to be a by-product of industrialized society. The glaciers of Europe are receding, and whether this could be the result of human activity or simply part of Earth's natural change as it comes out of the last ice age remains to be seen.

In the meantime, ice offers a great deal of potential for understanding our own planet and others. It may yet turn out that the ice sheets covering Mars contain single-cell life-forms. Furthermore, in August 1996, the National Aeronautics and Space Administration (NASA) reported that a meteorite found on the Antarctic ice may provide evidence of life on Mars. It seems that the 4.1-lb. (2 kg) meteorite contains polycyclic aromatic hydrocarbons that may have existed on that planet several billion years ago.

WHERE TO LEARN MORE

Alley, Richard B. *The Two-Mile Time Machine: Ice Cores, Abrupt Climate Change, and Our Future.* Princeton, NJ: Princeton University Press, 2000.

Bolles, Edmund Blair. *The Ice Finders: How a Poet, a Professor, and a Politician Discovered the Ice Age.* Washington, DC: Counterpoint, 1999.

Charlotte the Vermont Whale: Glaciers and the Glacial Ages (Web site). <http://www.uvm.edu/whale/GlaciersGlacialAges.html>.

Fagan, Brian M. *The Little Ice Age: How Climate Made History, 1300–1850.* New York: Basic Books, 2000.

Gallant, Roy A. *Glaciers.* New York: Franklin Watts, 1999.

"Home of Deep Glaciology." California Institute of Technology Division of Geological and Planetary Sciences (Web site). <http://skua.gps.caltech.edu/>.

Ice Ages (Web site). <http://www.museum.state.il.us/exhibits/ice_ages/>.

National Oceanic and Atmospheric Administration (NOAA) at the National Snow and Ice Data Center/World Data Center for Glaciology (Web site). <http://nsidc.org/noaa/>.

Stone, Richard. *Mammoth: The Resurrection of an Ice Age Giant.* Cambridge, MA: Perseus, 2001.

Virtually the Ice Age (Web site). <http://www.creswell-crags.org.uk/virtuallytheiceage/>.

METEOROLOGY AND THE ATMOSPHERE

EVAPOTRANSPIRATION AND PRECIPITATION

WEATHER

CLIMATE

EVAPOTRANSPIRATION AND PRECIPITATION

CONCEPT

Evaporation, along with the less well known process of transpiration, is the means by which water enters the atmosphere in the form of moisture. In evaporation liquid water from nonliving sources, such as the soil and surface waters, is converted into a gas. This conversion is driven by the power of the Sun, whose energy is also behind the process of transpiration, whereby plants lose water through their leaves. As with evaporation, transpiration places water in the atmosphere, and because the two processes work in tandem, they usually are spoken of together under the name evapotranspiration. Both make possible the formation of clouds, which, when they become saturated with moisture, produce the forms of precipitation by which water returns to the solid earth.

HOW IT WORKS

THE MOVEMENT OF WATER

The hydrosphere is the sum total of Earth's water, with the exception of water in the atmosphere. The hydrologic cycle is the continuous circulation between these two earth systems, hydrosphere and atmosphere, as well as the two other principal earth systems, biosphere and geosphere (see Earth Systems). Evapotranspiration and precipitation are principal components of this cycle, which is discussed in depth within the Hydrologic Cycle essay.

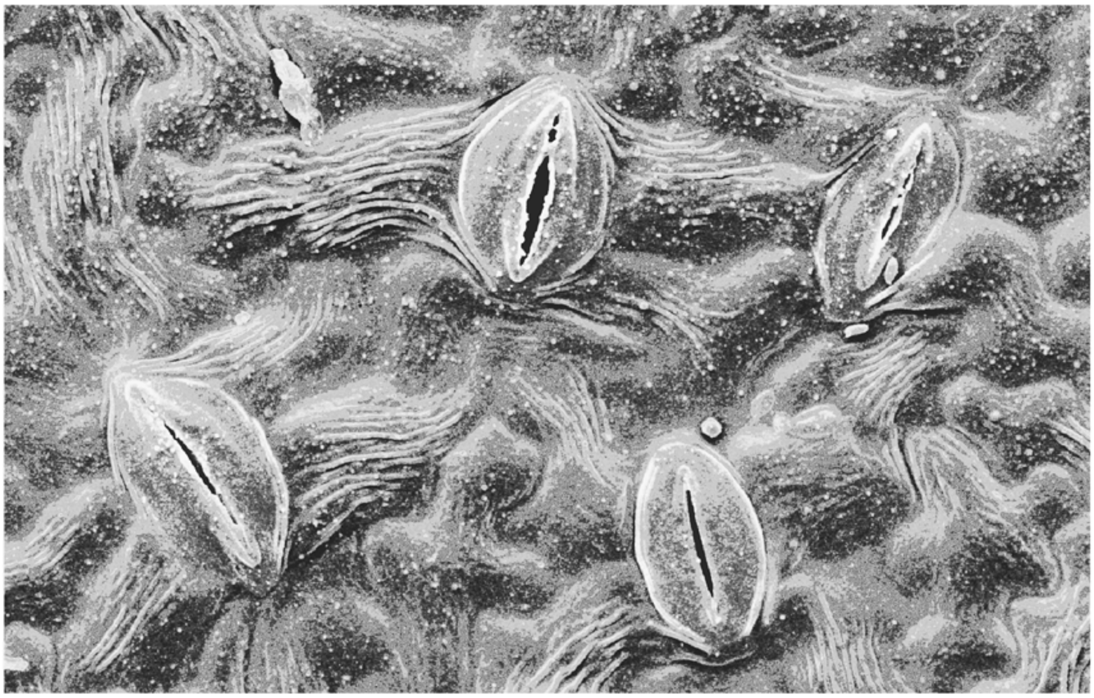
In evaporation, heat converts liquid water to a gaseous state, thus allowing it to be transferred to the atmosphere in the form of vapor. Whereas evaporation involves the loss of water to the atmosphere from nonliving sources, transpiration is the movement of water from living organisms to the atmosphere. This is achieved by the release of water through the plants' stomata, small openings on the undersides of leaves. In both cases, the Sun's electromagnetic energy, experienced as heat, is the driving mechanism, and because these phenomena are so closely related, they are normally treated together as evapotranspiration.

EARTH'S WATER BUDGET

Our present concern is primarily with the atmosphere; nonetheless, it is important to consider the other "compartments" in which water is stored. Some of them are regular way stations in the hydrologic cycle, but in the case of groundwater at least, the compartment is just that—a storage area from which it is conceivable that water might not be moved for millions of years.

The vast majority of water on Earth is indeed stored away, deep beneath the planet's surface in the form of groundwater in the lithosphere. This accounts for a staggering 94.7% of Earth's water. (For figures on the mass of this and other components in the water supply, see Hydrologic Cycles.) Much of the remainder is made up of the oceans, which account for 5.2% of the water on Earth, while glaciers and other forms of permanent and semipermanent ice make up 0.065%.

THE LAST 0.035%. We have now identified 99.965% of all water, yet we have not even approached any of the forms of water with which most of us typically come into contact. Of the remaining 0.035%, shallow groundwater, the source of most local water supplies, makes up the



SCANNING ELECTRON MICROGRAPH OF THE STOMATA, OR PORES, ON THE UNDERSIDE OF AN APPLE TREE LEAF. LEAF PORES PULL IN CARBON DIOXIDE FOR PHOTOSYNTHESIS. *(© Scimat/Photo Researchers. Reproduced by permission.)*

bulk, 0.30% of the total. Next are the inland surface waters. All of them combined, including such vast deposits as the Great Lakes and the Caspian Sea as well as the Mississippi-Missouri, Amazon, and Nile river systems—account for just 0.03% of Earth's water.

Now we are left with only 0.02% of the total, which is the proportion occupied by moisture in the atmosphere: clouds, mist, and fog as well as rain, sleet, snow, and hail. While it may seem astounding that atmospheric moisture is such a small portion of the total, this fact says more about the vast amounts of water on Earth than it does about the small amount in the atmosphere. That "small" amount, after all, weighs 1.433×10^{13} tons (1.3×10^{13} tonnes), or 28,659,540,000,000,000 lb.

WATER TURNOVER

The smaller the compartment of water, the greater the amount of turnover—that is, the exchange of "new" water for "old"—and the shorter the turnover time. Groundwater may stay put in aquifers, underground rock formations, for millions of years; on the other hand, atmospheric water experiences an enormous amount of turnover in just a year's time.

The atmosphere receives vast inputs of evaporation from the oceans as well as evapotranspiration from terrestrial ecosystems, or land-based communities of organisms. In the course of the year, the water in the atmosphere turns over about 34 times. Thus, the inputs of evaporation and transpiration are balanced almost perfectly by outputs of precipitation, which return more than 75% of atmospheric moisture to the oceans. The remainder falls on the land, where it contributes to brooks, streams, and rivers. The land receives 67% more water as precipitation than it loses through evaporation. The difference is made up by runoff to the oceans, primarily through rivers and in much smaller portions by subterranean channels.

HOW DOES THE WATER MOVE?

Clearly, water is moving through our atmosphere, but *how?* The answer is through gradients, or differences, of energy. When you hold a stone over the side of a cliff, preparing to drop it, the stone possesses a large quantity of gravitational potential energy. This potential energy is a function of the large gradient between the position of the stone and that of the ground.

In the same way, rivers and streams are driven by the gravitational potential that exists by virtue of the gradient from their source to the place where they empty into a lake or ocean. By definition, water flows downhill, and therefore the source of the river must be at a higher elevation than its delta, or the place where it discharges.

This matter of elevation is reflected in the confusing names for the two kingdoms that united in about 3100 B.C. to form Egypt: Upper Egypt was actually to the south of Lower Egypt. The names reflect the importance of the Nile, which flowed from the higher elevations of Upper Egypt into the fertile lowlands of Lower Egypt, home of Egyptian civilization's greatest farmlands.

EVAPORATION. Evaporation likewise is driven by energy gradients; in this case, however, the energy is not gravitational but electromagnetic. Specifically, it is the energy from the Sun, which we experience in the form of light and heat. (See Sun, Moon, and Earth for more about the Sun's energy.) As surfaces on Earth absorb solar electromagnetic energy, it increases their heat content and provides a source of energy to drive evaporation.

The second law of thermodynamics, discussed in Energy and Earth, tells us that the flow of heat is always from a high-temperature reservoir or area to a low-temperature one. Thus, if you hold a snowball in your hand, you may think that the snowball is cooling your hand, but, in fact, the opposite is happening: heat is passing from your hand into the snowball and warming it. As this happens, your hand loses heat, which you perceive as cold, though no coldness has been transferred—only heat.

In the same way, the energy difference between a heated surface and the atmosphere, manifested as a difference in temperature, makes it possible for water to be vaporized and transported into the air. Though the water has risen rather than fallen, its rise is brought about by the same principle that makes it possible for objects to fall from a great height, that is, a difference in potential. Thus it can be said that in evaporating, water "falls" from an area of high heat and high energy to one of low heat and low energy.

REAL-LIFE APPLICATIONS

Transpiration

When it comes to putting moisture into the atmosphere, transpiration is at least as important as evaporation. In fact, it puts more water into the air than evaporation does: any large area of vegetation tends to evaporate much larger quantities of moisture than an equivalent nonfoliated region, such as the surface of a lake or moist soil. Physically, however, the process of transpiration is the same as that of evaporation.

The only difference is that in the case of transpiration, the source of the evaporation is a biological one—leaves, for instance, as well as skin or lungs. From an environmental standpoint, the most important kind of transpiration is that which occurs through leaves. The loss of water from foliage puts an enormous amount of moisture into the atmosphere, and for this reason, areas where foliage appears in high concentration (i.e., forests) are vital to the cycling of water through the atmosphere.

Plants lose their water through moist membranes of a tissue known as *spongy mesophyll*, found in the tiny cavities that lie beneath the microscopic leaf pores called stomata. Stomata remain open most of the time, but when they need to be closed, guard cells around their borders push them shut. Because plants depend on them to "breathe" by pulling in carbon dioxide (see Carbon Cycle), however, they keep their stomata open—just as a human's pores must remain open, or the person would die of suffocation.

Because stomata are exposed in order to receive carbon dioxide for the plant's photosynthesis, this also means that the stomata are open to allow the loss of moisture to the atmosphere. It can be said, then, that transpiration—vital as it is to the functioning of our atmosphere—is actually an unavoidable consequence of photosynthesis, an unrelated process.

TRANSPIRATION AND ANIMALS. As we have suggested, transpiration in animals (including humans) takes place for much the same reason as it does with plants, as a by-product of breathing. Animals have to keep their moist respiratory surfaces, such as the lungs, open to the atmosphere. We may not think of our own breathing as transferring moisture to the air, but with just a little consideration of the subject it becomes clear that this is the case. The presence of moisture in our lungs can be proved simply by breathing on a piece of glass and observing the misty cloud that lingers there.

On the one hand, transpiration can cause animals to become dehydrated, but it also can be important in cooling down their bodies. When human bodies become overheated, they produce perspiration, which cools the surface of the skin somewhat. If the air around us is too humid, however, it already is saturated with water, and the perspiration has no place to evaporate. Therefore, instead of continuing to cool our bodies, the perspiration simply forms a sticky film on our skin. Assuming that the air is capable of absorbing more moisture, however, the sweat will evaporate, cooling our bodies considerably.

Heat, Cold, and Evapotranspiration

The preceding discussion brings up several more points about the relationship between heat and evapotranspiration. First of all, everything that is living has some degree of heat; if it did not, it would be at a temperature known as absolute zero, or 0K (–459.67°F or –273.15°C), which is impossible to reach. Therefore, even in Greenland there is a small amount of molecular motion in plant tissues, even when they appear to be completely frozen.

Naturally, however, there is very little evapotranspiration when temperatures are extremely cold, which is the reason why it can sometimes be "too cold to snow": temperatures are too low for sufficient moisture to be moved to the atmosphere. Nonetheless, there still can be some physical evaporation as the result of the direct vaporization of solid water, a process called sublimation.

HOT, DRY SUMMERS. At the height of summer, when air temperatures are warm and the trees are fully foliated (i.e., covered in leaves), a high rate of transpiration occurs. So much water is pumped into the atmosphere through foliage that the rate of evapotranspiration typically exceeds the input of water to the local environment through rainfall. The result is that soil becomes dry, some streams cease to flow, and by late summer, in extremely warm temperate areas such as the southern United States, there is a great threat of drought and related problems, such as forest fires.

Once the trees drop their leaves in the autumn, transpiration rates decline greatly. This makes it possible for the parched soil to become recharged by rainfall and for streams to flow again. Such is the case in a temperate region, which by definition is one that has the four seasons to which most people in the United States (outside Hawaii, Alaska, and extreme southern Florida and Texas) are accustomed. In a tropical region, by contrast, there is a "dry season," in which transpiration takes place, and a "rainy season," in which moisture from the atmosphere inundates the solid earth. This rainy season may be so intense that it produces floods.

Clouds

So far, we have talked mostly about the means by which water moves from the solid earth into the atmosphere. Now let us consider what happens when it gets there, at which point—assuming there is sufficient moisture—it will coalesce to form a cloud. Though most people are inclined to romanticize clouds, seeing in them shapes and colors and sometimes even reflections of their own moods, clouds are nothing more than atmospheric moisture that has condensed to form tiny droplets of water or crystals of ice.

Heated, moist air that rises from the ground is much more dense than the air that lies above it, but as it rises, it expands and becomes less dense. This expansion cools the air, so that the water vapor condenses into tiny droplets. As a result, a cloud forms, and the relative intensity of the energy gradients that brought about the formation of the cloud creates different shapes.

For example, if there is a vigorous uplift of air, resulting from sharp differences in temperature and pressure between the ground and the atmosphere, the resulting clouds will have a tall, stacked appearance. On the other hand, clouds formed by the gentle uplift of air currents have a flat, cottony appearance. Thus, the shapes of the clouds themselves, as well as the ways in which they change, assist meteorologists in predicting the weather.

Cloud Classifications

Thanks to a system developed in 1803 by the English pharmacist and amateur naturalist Luke Howard (1772–1864), it is possible to classify clouds according to three basic shapes. These shapes are known by the Latin names *cumulus* (piled heaps and puffs), *cirrus* (curly, fibrous shapes), and *stratus* (stretched and layered).

Howard combined these names with adjectival terms, such as *alto* ("high") and *nimbus* ("rain"), to describe variations on the three basic cloud types. Today, the International Cloud Classification used by meteorologists worldwide applies Howard's system to the identification of ten basic cloud types, or genera. They are divided into three high-altitude genera, two midlevel ones, three low-level genera, and two varieties of rain cloud.

HIGH-LEVEL CLOUDS. The three genera of high-level clouds appear at a range between 16,500 ft. and 45,000 ft. (5,032–13,725 m), though they usually form in a belt between 20,000 ft. and 25,000 ft. (6,000–7,500 m). All are cirrus clouds, of which the highest are pure cirrus, made of ice crystals.

Then there are cirrocumulus, the least common variety of cloud. Composed of either supercooled water droplets (i.e., water that continues to exist in liquid form even though the temperature has dropped below the freezing point) or ice crystals, these are small and white or pale gray, with a rippled appearance like oatmeal. Finally, cirrostratus clouds—which, like pure cirrus clouds, are made of ice crystals—often form a halo around the Sun or Moon in the wintertime.

MIDLEVEL CLOUDS. The two genera of midlevel clouds, which appear at 6,500-23,000 ft. (2,000–7,000 m), are named with the prefix "alto." Altostratus clouds are usually a uniform bluish or gray sheet covering large portions of the sky and either completely or nearly obscure the Sun and Moon. These complex clouds are composed of layers, with ice crystals at the top, ice and snow in the middle, and water droplets in the lower layers.

Altocumulus are oval-shaped, dense, fluffy balls that may appear either as singular units or as closely bunched groups. When sunlight or moonlight shines through these clouds, the light often is perceived from the ground in the form of rays.

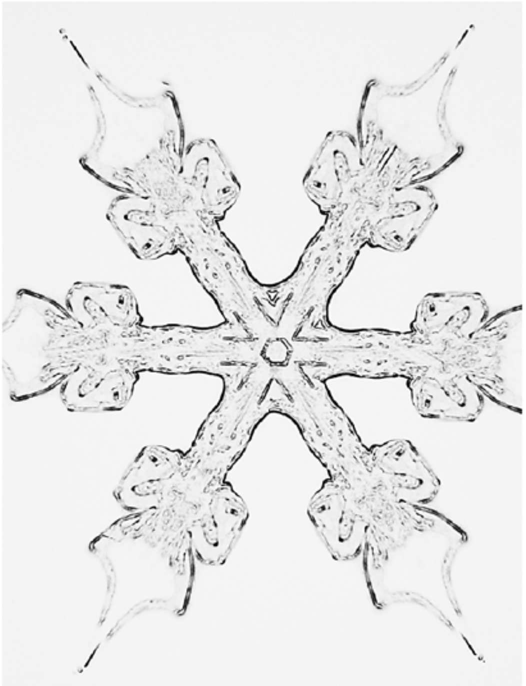
LOW-LEVEL CLOUDS. There are three genera in the low level, which extends from the surface to 6,500 ft. (2,000 m). Pure stratus clouds, which are generally the lowest, blanket the sky and typically appear gray. They are formed when a large mass of air rises slowly or when cool air moves in over an area close to ground level. Stratus clouds may produce mist or drizzle, or, when they form at ground level, they may appear as fog.

When moisture coalesces in the atmosphere, it forms a cloud. Altocumulus clouds, which look like oval-shaped, dense, puffy balls, are midlevel clouds. *(© Mark A. Schneider/Photo Researchers. Reproduced by permission.)*

Pure cumulus clouds are flat on the base and vertically thick, with a puffy appearance on top. This puffiness is the result of updrafts. Occurring primarily in warm weather, cumulus clouds are made of water droplets and look brilliant white because of the sunlight's reflection off the droplets. Last, stratocumulus clouds are large, grayish, and puffy, often looking like dark rolls.

RAIN CLOUDS. Stratocumulus clouds can transform into nimbostratus clouds, while cumulus can develop into cumulonimbus clouds. These two—nimbostratus and cumulonimbus—are the last of the ten varieties of cloud, designated not by altitude but by the fact that they give rise to precipitation. (Note that the light, airy variety known as *cirrus* never portends rain.)

Nimbostratus usually appear at midlevel altitudes. Made of water droplets that produce either rain or snow, they are thick, dark, and gray.



Micrograph of a snowflake. Snowflakes are formed by the bonding of solid ice crystals inside a cloud. *(© Kent Wood/Photo Researchers. Reproduced by permission.)*

Sometimes they are seen with virga, which are skirts of rain trailing down their sides.

Cumulonimbus, or thunderstorm, clouds arise from cumulus clouds that have reached a great height. At a certain height, the cloud flattens out, and powerful updrafts and downdrafts create a great deal of unrest inside the cloud, which contains all phases of water: gas, liquid, and solid. These clouds can cause violent storms.

Rain, Snow, Hail, and Sleet

Eventually, the amount of moisture in the cloud becomes so great that it has to fall earthward in the form of precipitation—usually rain, snow, hail, or sleet. These types of precipitation are distinguished by the form they take: liquid in the first case, lightly frozen particles in the second case, and hard, frozen nuggets in the latter two cases.

Liquid precipitation includes drizzle and raindrops, which are distinguished from each other on the basis of size. Raindrops have a radius of about 0.04 in. (1 mm), while drops of drizzle are about one-tenth that size. Note the use of the term radius, implying a sphere. Drops of liquid precipitation are spherical, and though a teardrop shape is popularly associated with raindrops, they assume that shape only when they are falling.

Snowflakes are formed by the aggregation, or physical bonding, of solid ice crystals inside a cloud, while hailstones are made of a combination of supercooled water droplets and pellets of ice. Not only are they more dense than snowflakes, but they are also more spherical. Similar to hail is sleet, pellets of pure ice that usually are much smaller than hail.

FORMING PRECIPITATION. The type of precipitation formed depends on the warmth of the cloud from which it comes. "Warm" clouds are those whose temperatures are above freezing—32°F (0°C)—and "cold" clouds are those that are at least partially below freezing temperature. These temperature values are themselves a function of altitude, since temperature in the lower atmosphere (where all precipitation occurs) decreases by about 5.32°F for every mile (6°C per kilometer) of altitude gained.

As we have suggested earlier, humidity relates to the ability of air to evaporate, and, therefore, when the air has all the water it can hold, it is said to have a relative humidity of 100%. The cooler air is, the less moisture is required for it to become saturated. Once saturation occurs, moisture molecules form around cloud condensation nuclei, such as nitrates or extremely fine particles of sea salt that have managed to evaporate.

Cold clouds form around ice crystals. Before freezing, the water in these clouds may be supercooled,. Ice nuclei are much more rare than cloud condensation nuclei and are less well understood by meteorologists.

MIXED CLOUDS. Clouds that contain both liquid water and ice are called mixed clouds. Supercooled water will freeze when it strikes an ice crystal, and if it freezes immediately, it forms what is known as opaque or rime ice. If it freezes slowly, the result is called clear ice. Eventually the ice forms a thick coating, which is the origin of hail.

Not all mixed clouds produce frozen precipitation, however. Thunderstorms, involving electric charges imparted to precipitation particles, which leads to the eventual discharge of lightning, are also the products of mixed clouds. (For more about storms and precipitation, see Weather and Climate.)

KEY TERMS

AQUIFER: An underground rock formation in which groundwater is stored.

ATMOSPHERE: In general, an atmosphere is a blanket of gases surrounding a planet. Unless otherwise identified, however, the term refers to the atmosphere of Earth, which consists of nitrogen (78%), oxygen (21%), argon (0.93%), and other substances that include water vapor, carbon dioxide, ozone, and noble gases such as neon (0.07%).

BIOSPHERE: A combination of all living things on Earth—plants, mammals, birds, amphibians, reptiles, aquatic life, insects, viruses, single-cell organisms, and so on—as well as all formerly living things that have not yet decomposed.

ECOSYSTEM: A community of interdependent organisms along with the inorganic components of their environment.

EVAPORATION: The process whereby liquid water is converted into a gaseous state and transported to the atmosphere. When discussing the atmosphere and precipitation, usually evaporation is distinguished from transpiration. In this context, evaporation refers solely to the transfer of water from nonliving sources, such as the soil or the surface of a lake.

EVAPOTRANSPIRATION: The loss of water to the atmosphere via the combined (and related) processes of evaporation and transpiration.

GEOSPHERE: The upper part of Earth's continental crust, or that portion of the solid earth on which human beings live and which provides them with most of their food and natural resources.

GROUNDWATER: Underground water resources that occupy the pores in bedrock.

HYDROLOGIC CYCLE: The continuous circulation of water throughout Earth and between various earth systems.

HYDROLOGY: The study of the hydrosphere, including the distribution of water on Earth, its circulation through the hydrologic cycle, the physical and chemical properties of water, and the interaction between the hydrosphere and other earth systems.

HYDROSPHERE: The entirety of Earth's water, excluding water vapor in the atmosphere but including all oceans, lakes, streams, groundwater, snow, and ice.

METEOROLOGY: The study of the atmosphere, weather, and weather prediction.

POTENTIAL: Position in a field, such as a gravitational force field.

POTENTIAL ENERGY: The energy that an object possesses by virtue of its position or its ability to perform work.

PRECIPITATION: When discussing the hydrologic cycle or meteorology, precipitation refers to the water, in liquid or solid form, that falls to the ground when the atmosphere has become saturated with moisture. In the context of chemistry, precipitation refers to the formation of a solid from a liquid.

SUPERCOOLED: A term for water that continues to exist in liquid form even though its temperature has dropped below the freezing point.

KEY TERMS CONTINUED

SYSTEM: Any set of interactions that can be set apart mentally from the rest of the universe for the purposes of study, observation, and measurement.

TRANSPIRATION: The process whereby moisture is transferred from living organisms to the atmosphere. A major portion of environmental moisture for precipitation comes from plants, which lose water through their stomata, small openings on the undersides of leaves.

WATERSHED: An area of terrain from which water flows into a stream, river, lake, or other large body.

WHERE TO LEARN MORE

Bad Meteorology (Web site). <http://www.ems.psu.edu/~fraser/BadMeteorology.html>.

Global Hydrology and Climate Center (Web site). <http://wwwghcc.msfc.nasa.gov/>.

Glover, David. *Flying and Floating.* New York: Kingfisher Books, 1993.

ISCCP: International Satellite Cloud Climatology Project (Web site). <http://isccp.giss.nasa.gov/isccpframe.html>.

Kahl, Jonathan D. *Wet Weather: Rain Showers and Snowfall.* Minneapolis: Lerner Publishing, 1992.

Parker, Steve. *Weather*. Illus. Kuo Kang Chen and Peter Bull. New York: Warwick Press, 1990.

Schmid, Eleonore. *The Water's Journey.* New York: North-South Books, 1990.

Tropical Rainfall Measuring Mission/TRMM (Web site). <http://www-tsdis.gsfc.nasa.gov/tsdis/tsdis.html>.

The World-Wide Web Virtual Library: Meteorology (Web site). <http://dao.gsfc.nasa.gov/DAO_people/towens/VLM/>.

Wyman, Richard L. *Global Climate Change and Life on Earth.* New York: Routledge, Chapman and Hall, 1991.

WEATHER

CONCEPT

Though people are quite accustomed to experiencing weather—sunshine and rain, wind and storms, hot and cold, fair days and foul—most have little idea how weather originates or even really what weather is. In fact, it is a condition of the atmosphere produced by one or more of six factors: air temperature, air pressure, humidity in the air, the density and variety of cloud cover, the amount and type of precipitation, and the speed and direction of the wind. Those six variables can produce the sunny, beautiful days that everyone treasures, or they can conspire to create the sort of rainy, windy, and cold day that some people despise and others (especially creative types, who do not have to be out in the elements) adore. The variables of weather also can align to manifest in the form of brutal storms, including hurricanes and tornadoes. Yet as complex as weather is, its behavior can still be forecast—with at least some accuracy—and sometimes even manipulated.

HOW IT WORKS

Introduction to Weather

Weather and climate are not the same thing: weather is the condition of the atmosphere at any given time and place, while climate is the pattern of weather conditions in a particular region over an extended period. We do not speak therefore of "forecasting climate," though it is possible, on the basis of long-term climate patterns, to make some predictions concerning future climatic trends. (Climate is discussed in a separate essay.)

Nevertheless, the type of atmospheric prediction that applies much more to our daily lives is weather forecasting. The latter is the work of meteorologists, or scientists involved in studying the atmosphere and weather and in making weather predictions. In the 1970s most local television news programs still included a "weatherman" on their lineup. This was a television personality, much like the anchors and the specialists in sports or other topics, whose job it was to report the weather. During the late 1970s and early 1980s, however, most stations and networks changed their designation of weatherman to that of meteorologist.

In part this change had to do with gender politics, since the title weatherman seemed sexist but weatherwoman or weatherperson sounded a bit absurd. On the other hand, it also was made in recognition of the work that weather forecasters perform: unlike ordinary reporters, they are not so much journalists as they are scientists. (Though a nice smile, charming personality, and good looks never hurt any TV meteorologist!) And while a meteorologist, like every member of the TV news team, reports what has happened—for instance, a record snowfall—his or her job also is concerned heavily with explaining what *will* happen.

SIX FACTORS. As we have noted, weather is determined by six factors, all of which involve the atmosphere, and several of which also relate to water in the atmosphere. These factors are as follows:

- Air temperature
- Air pressure
- Wind
- Humidity

- Cloud cover
- Precipitation

The role of air temperature in weather is fairly obvious, since that is one factor of which we are used to taking note on a regular basis. Air pressure, though much less familiar to most of us, has a powerful influence on air currents, and air pressure is linked closely with wind. Likewise, there are clear linkages among humidity (moisture in the air), the amount and type of cloud cover (formed, of course, from moisture in the air), and the amount and type of precipitation, which begins to fall when the clouds are saturated with moisture.

WHAT MAKES THE WEATHER?

So much for the factors involved in weather, which go into making the weather we experience at a given time. But if we want to search for the ultimate causes behind the weather, we have to look toward larger phenomena: the Sun and Earth as well as Earth's surface and atmosphere and the interactions among all of these factors.

At the heart of weather is the Sun and the energy it transmits to Earth. As discussed in Energy and Earth, the Sun sends out vast quantities of energy, only a small portion of which comes anywhere near Earth—and a much smaller portion reaches the surface of the planet as usable energy. Yet that "tiny" amount, which is the vast majority of the energy available to Earth in any form, is enough to light and warm the planet and to drive a variety of processes in the biosphere and atmosphere.

MOVING WATER THROUGH THE ATMOSPHERE. The Sun's most direct and obvious influence on weather is its impact on temperature, but this is only one aspect of a larger and more complex picture. As discussed in the essay Evapotranspiration and Precipitation, differences in temperature—a result, primarily, of the Sun's heat—make possible the conditions for both evaporation and transpiration, or the loss of moisture from living organisms to the atmosphere.

The amount of water that evaporates into the atmosphere in a given area is its humidity, which exerts a powerful impact on the human experience of weather. As most people who have traveled widely will attest, high temperatures in a dry climate, such as that of the American Southwest, are far easier to endure than equally high temperatures (or even lower ones) in Florida or other parts of the Deep South, a much more humid region.

As water moves from Earth's surface into the atmosphere through evapotranspiration (a combination of evaporation and the related process of transpiration), the air through which it passes becomes increasingly cooler. Eventually, the vaporized water comes to an altitude at which the air around it is cold enough to cause condensation—that is, the formation of a liquid. When enough vapor has condensed into tiny droplets of water, or tiny ice crystals, the result is the formation of clouds.

Clouds, discussed in detail within the context of Evapotranspiration and Precipitation, are integral to the formation of weather patterns, which is why TV meteorologists almost always show their audiences satellite maps illustrating cloud movements. Not only do clouds reflect sunlight into space, meaning that a sufficient accumulation of cloud cover will bring about a decrease in temperature, but clouds also manufacture precipitation. Forms of precipitation include rain in all its variations, from a fine mist to a downpour, as well as snow, sleet, and hail.

AIR PRESSURE AND WIND

We have seen how the Sun affects the movement of water through the atmosphere, thus driving three of the six key weather factors we named earlier: humidity, cloud cover, and precipitation. In addition to its effect on these factors and on temperature itself, the Sun and its energy are also behind the other two important factors—air pressure and the movement of air through the atmosphere, that is, winds.

Because it does not change as frequently or as dramatically as temperature (fortunately!), air pressure is something of which we are hardly aware. But if a person accustomed to the sea-level air pressure of a place such as New York City suddenly had to travel to a high-altitude locale such as Denver, Colorado, he or she immediately would become aware of the difficulty in breathing and the other changes that attend a change in pressure.

For instance, the higher the altitude (and, hence, the lower the air pressure), the lower the temperature at which water boils. For this reason, cake mixes and similar products have special instructions for kitchens at high altitudes. Taken to an extreme, this means that with absolutely no pressure, liquid could boil—that is, turn into a vapor—at extremely low temperatures. This is one of the reasons, along with lack of oxygen and the presence of harmful rays, that an astronaut wears a protective suit; without it, the pressureless environment of space would cause a person's blood to boil!

In the English system, normal atmospheric pressure at sea level is 14.7 lb. per sq. in., or 1.013×10^5 pascals (Pa) in the metric or SI (Scientific International) system. This amount of pressure constitutes a unit in its own right, an atmosphere (atm). Atmospheric pressure, however, usually is measured in terms of the bar, an SI unit equal to 10^5 Pa. Thus, 1 atm = 1.013 bars. Meteorologists often use the millibar (mb), which, as its name implies, is equal to 0.001 bars—roughly 1/1,000 of ordinary air pressure at sea level.

WIND: FROM HIGH PRESSURE TO LOW. Heat from the Sun does not fall evenly on all places on Earth. Aside from the obvious fact that it makes a limited impact on polar regions, there are the differences in color and texture between various areas, even in the most tropical latitudes—that is, those closest to the direct path of sunlight. For example, soil, since it is almost always darker than water, tends to attract more sunlight than bodies of water do.

When an area is heated, the air above it heats up as well, and this makes that region of air less dense. As a result, the air rises, a phenomenon known as convection. Convection currents also carry other masses of air downward from the upper atmosphere toward Earth's surface. In regions where warm air moves upward, the atmospheric pressure tends to be low, whereas downward air movements are associated with higher atmospheric pressures.

These higher or lower atmospheric pressures can be measured by a barometer, an instrument that registers pressure just as a thermometer measures temperature. The fact that there are differences in pressure between areas brings about wind, which is the movement of air from a region of high pressure to one of lower pressure. (Later in this essay, we discuss a way to "create" wind and thus test this statement.)

OTHER INFLUENCES ON WEATHER

We have focused primarily on the influence that the Sun and its energy exerts on weather, but it should be noted that certain aspects of Earth itself determine weather patterns. Among these factors, as noted earlier, are color and texture, whereby certain places on the planet are more apt to receive the Sun's energy than others.

Also important is Earth's position in space. First of all, there is the tilt of its axis relative to the plane of its rotation around the Sun, which accounts for the seasons. In addition, the planet follows an elliptical, or oval-shaped, orbital pattern around the Sun, which means that there are certain times when the planet is closer to the Sun and its energy than at others.

Earth's rotation brings about complicated patterns of movement on the part of air masses heated at the equator. If the planet were not rotating, these masses of air would simply move from the equator toward the poles, where they would be cooled. Because the planet is rotating, however, the movement of global winds is much more complex and is characterized by circular patterns known as *cells.*

In addition to these factors that relate to the planet's movement in space, there is also the matter of irregularities on Earth's surface. An example of this is the effect mountains can have on cloud movements, creating a perpetually rainy climate on one side and a dry "rain shadow" on the other. Rain shadows are discussed in the essay on Mountains.

REAL-LIFE APPLICATIONS

CREATING WIND

Earlier we discussed the fact that wind results from differences in pressure. This is a statement you can test for yourself—and perhaps you already have, without knowing it. Suppose you are in a room where the heat is on too high and there is no way to adjust the thermostat. Outside the air is cold, so you open a window, hoping to cool down the room. Does it do the trick? Not likely. But if you open the door leading from the room to the hallway, a nice cool breeze will blow through. You have, in effect, created wind or at least manipulated the conditions to make wind possible.

With the door closed, the room constitutes an area of high pressure in comparison with the area outside the window. Once the window alone is opened, it is theoretically possible for air to flow into the room, but that does not mean the air actually will flow. The reason is that fluids such as air—whether in a room or in the sky—tend to move from areas of high pressure to areas of low pressure. Because the room is of relatively high pressure compared with the outside, there is no reason for the air to move into the room.

Furthermore, in line with the second law of thermodynamics (see Energy and Earth), the flow of heat always will be from an area of relatively high temperature to one of relatively low temperature. Therefore, if there is going to be any air movement in this situation, it will have to be movement of hot air *out* of the room rather than cool air *into* the room. (Even an air conditioner works by taking out the heat, not by bringing in "cold"—even though we may perceive it otherwise.)

In the case of the overheated room, there is only one way to solve the problem: by opening the door into the hallway, or whatever else lies outside the door. As soon as the door is opened, the relatively high-pressure air of the room flows into the relatively low-pressure area of the hallway, just as the laws of physics say it should. This is exactly the same principle whereby wind blows from high-pressure regions into low pressure ones. By setting up a cross-draft, in effect, we have "created" wind.

Thunderstorms and Worse

Wind and clouds combined with rain (discussed, like clouds, in Evapotranspiration and Precipitation) and a few other factors create a thunderstorm. Those other factors, of course, are lightning and that key difference between an ordinary storm and a thunderstorm: thunder.

Inside a thunderstorm are updrafts and downdrafts of air, which bring about a buildup of static electric charges within the thunderstorm clouds. Over time, there is a large buildup of separate electric charges inside the cloud, with positive charges near the top and negative charges in the middle. This separation of charges creates huge voltage differences, which require something to equalize them: a bolt of lightning, which suddenly passes between the areas of differing charge.

Lightning produces a spark, which heats the air in an instant to more than 54,000°F (30,000°C). As a result, the molecules of air and moisture in the cloud experience an extremely sudden expansion, and this expansion is accompanied by a release of energy. The energy in this situation takes the form of sound, which we experience as thunder.

ANATOMY OF A THUNDERSTORM. Let us go back now to the point when the thunderstorm came into being. First solar energy, or some other influence, such as the presence of a mountain range, causes water to enter the atmosphere as vapor. As this air rises, it expands and cools, eventually condensing and coalescing to form a cloud. The cloud thus formed is a cumulus cloud, which may appear as the puffy, benign cloud of a clear summer's day or may turn into a cumulonimbus—a thunderstorm—cloud.

On its way to becoming a thunder cloud, a cumulus undergoes a transformation into a convective cloud, or one formed by convection. It may never become a thunder cloud at all, assuming that vertical movement stops, in which case fair weather prevails. But if the atmosphere is unstable, meaning that the air temperature drops rapidly with altitude, packets of air that begin rising and cooling inevitably will be warmer than the air around them. The rising air packet is like a balloon, weighing less per unit of volume than the surrounding area, and thus it continues to rise.

As we have noted, in order to evaporate, water needs to receive a certain infusion of energy, or heat, from the Sun. Once it condenses and forms a cloud, it releases that heat, warming the cloud and causing it to rise still higher. As long as these updrafts can support the cloud, it continues to grow and to rise, until, in the case of a thunderstorm, it reaches very great atmospheric heights—about 40,000 ft. (12 km) above the surface. In the course of rising and growing, large raindrops and hailstones can form.

TWO TYPES OF THUNDERSTORM. Over western Florida and other areas around the Gulf of Mexico, thunderstorms grow as we have described, with cumulus clouds rising and cooling. These are called air-mass storms. Once the cloud reaches a certain point of moisture saturation, rain begins to fall from the upper part of the cloud, producing precipitation and with it downdrafts. Because the downdraft in such a situation is usually stronger than the updraft, the cloud is likely to dissipate before the thunderstorm has caused much damage. Certainly there will be rain showers, thunder, and lightning, but the storm is unlikely to produce extreme wind damage or hail.

On the other hand, parts of the central and eastern United States are prone to what are called frontal thunderstorms, and these storms are much more severe. They typically form just ahead of a cold air mass, or cold front. The cold air, much more dense than the warm, humid, and unstable air of the cloud system, pushes the clouds ahead of it, which rise and form convective clouds. Eventually, these clouds produce rain, which causes downdrafts. But instead of dissipating the storm's impact (as in the case of the air-mass storm), the downdrafts only increase its intensity.

In a frontal thunderstorm, strong downdrafts hit the ground, spreading out and sending more warm humid air rising into the storm. This gives the storm clouds more latent heat, increasing the updrafts, which in turn gives more speed to the wind and improves the chances of heavy rain or even hail. The latter can appear even in summertime or in warm climates, since the cloud forms at a great height.

When ice crystals form in the cloud, the updrafts and downdrafts circulate them back and forth, and as this happens, the crystals pick up more and more moisture. Eventually, tiny ice crystals gather rainwater around them, until they have become hailstones. Depending on the conditions of the storm, hailstones can become extremely large—as big as 5.5 in. (14 cm) in diameter.

Tornadoes

Some forms of weather make thunderstorms seem minor by comparison. Among these are tornadoes, a rapidly spinning column of air formed in a severe thunderstorm. The vortex, or rotating center, of the column, forms inside the storm cloud and begins to grow downward until it touches Earth's surface. The United States is particularly prone to tornadoes, owing to specific factors of its location and its larger climate patterns. (See Climate for more about the distinction between weather and climate.)

The type of severe frontal thunderstorm that can produce a tornado typically is associated with an extremely unstable atmosphere and with moving systems of low pressure, which bring

As clouds rise and grow, ice crystals form and gather rainwater around them until they become hailstones. Depending on the conditions of the storm, hailstones can grow extremely large—more than twice the size of the golf ball–size ones shown here. *(© Gary Meszaros/Photo Researchers. Reproduced by permission.)*

masses of cold air into contact with warmer, more humid air masses. It so happens that such storms occur frequently across a wide swath of North America, from the plains states to the eastern seaboard (i.e., more or less from Kansas to North Carolina) during a period from about June to October each year.

As updrafts in a severe thunderstorm cloud become stronger, they pull more air into the base of the cloud to replace the rising air. As the air from the surface moves into a smaller area, it begins to rotate faster, owing to what physicists call the conservation of angular momentum. The latter can be illustrated by the example of an ice skater who, while spinning with her arms outstretched, pulls her arms inward. As she does so, the speed of her rotation increases. The same happens with rotating air as it moves from the large space of the ground to the smaller space of the cloud.

Thus, a funnel cloud is produced and grows, and if it becomes large enough, it may touch ground—with devastating results. Within the vortex of a tornado, which is typically about 328 ft. (100 m) in diameter, wind speeds may be greater than 220 MPH (100 m/s). The tornado itself travels at speeds of 10–30 MPH (15–45 km/h), making a sound like a freight train or a jet engine and wreaking havoc along a path as long as 200 mi. (321 km). Nothing can stand in the way of a tornado with enough force: buildings shatter, roofs and whole houses take to the air, and pieces of straw can be blown through solid wood.

TORNADO ALLEY. No country in the world is more prone to tornado activity than the United States, a large portion of which is known as "Tornado Alley." The latter area has no specific boundaries, though it is more or less contiguous with the wide Kansas-to-Carolina swath mentioned earlier. Some authorities describe Tornado Alley as including northern Texas, Oklahoma, Kansas, and southern Nebraska. However, the American Meteorological Society's *Glossary of Weather and Climate* defines Tornado Alley as "The area of the United States in which tornadoes are most frequent. It encompasses the great lowland areas of the Mississippi, the Ohio, and lower Missouri River valleys. Although no state is entirely free of tornadoes, they are most frequent in the plains area between the Rocky Mountains and the Appalachians."

It is no accident that a wide, flat region, over which heavy winds can blow from numerous directions—including winds off of the Gulf of Mexico and Atlantic Ocean—would be home to such enormous tornado activity. Tornado Alley, or parts of it, has seen numerous extraordinary weather events in which not one or two tornadoes struck, but many dozens—a phenomenon known as a tornado outbreak, or a super outbreak.

The worst super outbreak in recent memory occurred on April 3–4, 1974, when 148 rampaged through 14 states. It began on the morning of the 3rd, when a low-pressure system moved over central Kansas, spreading a cold front as far south as Texas. Meanwhile, a warm front clung to the lower Ohio River Valley, and various unstable patterns covered the South. Into this motley mix came strong winds, which soon took hold over much of the eastern United States. By afternoon, thunderstorms began raging over much of the Midwest and South, including the Ohio and Tennessee River valleys.

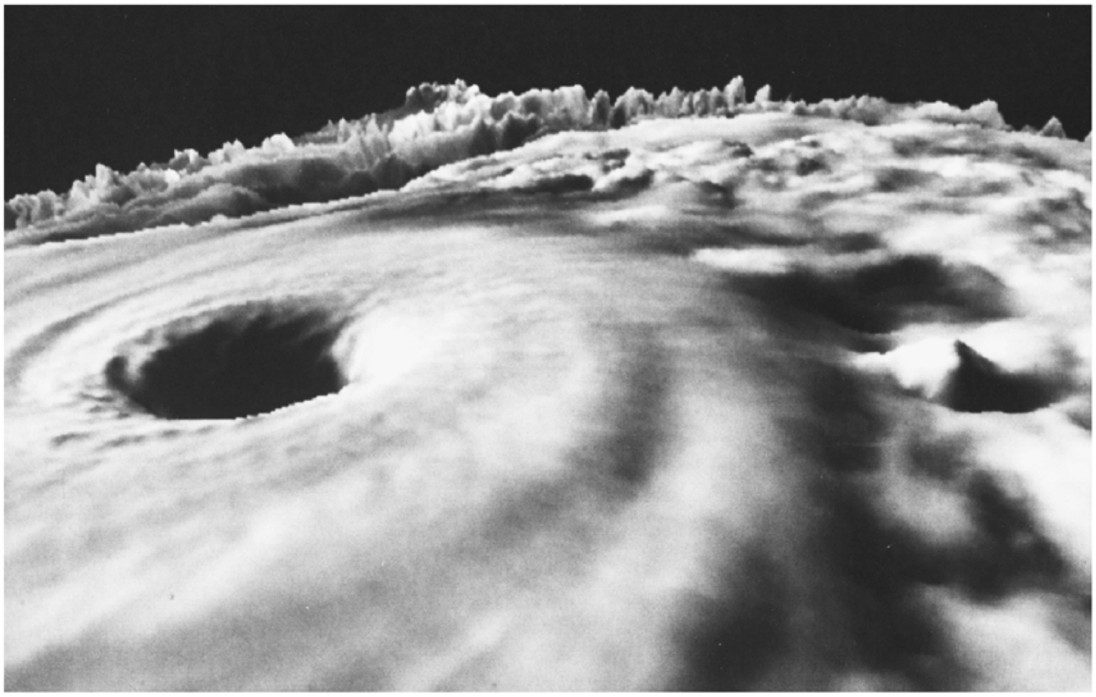
The result was a series of nearly 150 tornadoes, including 48 that were classified as F4 or F5, the highest levels on the Fujita-Pearson scale used to rate the intensity or tornadoes. Some other significant tornado events since that time include the Carolina outbreak, which killed 57 people, injured 1,248, and caused $200 million in damage, on March 28, 1984. A year later, on May 31, 1985, an outbreak of 41 tornadoes in Ohio, Pennsylvania, New York, and Ontario killed almost 90 people and caused over $450 million in damage.

CYCLONES

Another form of extreme weather is a cyclone, a general term for what is sometimes called a hurricane or a typhoon. These are vast circulating storm systems characterized by bands of showers, thunderstorms, and winds. They develop over warm tropical oceans, generally as isolated thunderstorms, and turn into monsters. A cyclone may have a diameter as great as 403 mi. (650 km) and be more than 7.5 mi. (12 km) in height.

Near the center of the cyclone, winds may be as high as 110 MPH (50 m/s), yet the very center is an area of complete calm, known as the eye. This is a region of descending air surrounded by the updrafts that characterize the cyclone, making the eye a little funnel of peace in the middle of a terrible storm. If it were possible to stay in the eye of the cyclone as it moves—at speeds comparable to that of a tornado—it is conceivable one could come away unscathed.

THE IMPACT OF CYCLONES. Once the cyclone blows inland and reaches population centers, it can cause massive death and destruction. The southern United States witnessed this with hurricanes such as Hugo, which devastated the Carolinas and nearby regions in 1989. The damage on Puerto Rico alone (one of the places Hugo touched down before moving north) amounted to 12 lives lost, and $2 billion in property. Thanks to evacuation measures and a bit of good fortune (the hurricane missed Charleston, South Carolina), loss of life was minimal on the continental United States, and loss of property was kept to $5 billion. By contrast, Hurricane Andrew, which struck Florida and other parts of the southern United States in August



A GLOBAL NETWORK OF WEATHER SATELLITES MAKES IT POSSIBLE TO IDENTIFY AND TRACK TROPICAL CYCLONES FROM THEIR EARLIEST APPEARANCE AS DISTURBANCES OVER THE OCEAN. A CYCLONE CAN GENERATE WINDS AS HIGH AS 110 MPH (50 M/SEC), YET THE VERY CENTER IS AN AREA OF COMPLETE CALM, KNOWN AS THE EYE. *(U.S. National Aeronautics and Space Administration [NASA].)*

1992, was the costliest natural disaster in U.S. history, with over $25 billion in damage reported.

But the impact of storms in the United States is dwarfed by the destruction of hurricanes and typhoons in third world countries. Bangladesh, for instance, is a country where a population half that of the United States is crammed into an area the size of Wisconsin. (And, it should be noted, a country so poor that the income of the average full-time working man is far less than that of an American teenager working a part-time job.)

In such a situation, the impact of cyclones is bound to be devastating. Though dollar figures of property damage from the country's many cyclones are not available, the human toll is better known. Since 1963, when it was still called East Pakistan, Bangladesh, has seen seven cyclones in which 10,000 or more died. The worst, in 1970, killed half a million. Government estimates of deaths from a cyclone on April 10, 1991, were 150,000, though it is likely that many more died from disease, starvation, and exposure.

Likewise, typhoons in such countries as the Philippines cause massive power outages, flooding, landslides, and destruction of life and property. In part this is because a number of these countries are located in or near tropical zones that are especially susceptible to cyclones, but it is also a matter of preparedness.

RESPONDING TO CYCLONES. In the United States, with its greater material wealth and technological sophistication, it is possible to prepare people for extreme weather in a way that is simply beyond the reach of many less prosperous or powerful nations. It is not so much that Americans are capable of developing structures such as seawalls that can withstand the force of hurricanes, though this certainly helps. The seawall in Charleston, South Carolina, for instance—a massive bulwark of concrete more than 10 ft. (3 m) high—is intended to protect homes in that city's historic Battery when hurricanes produce powerful ocean swells.

Much of the effectiveness of the American response to hurricanes, however, is attributable to the ability to react to circumstances rather than to prevent them. With a large amount of the population possessing electronic communication, it is possible to circulate word quickly regarding the coming deluge. Furthermore, people in the United States are much more mobile

DRY ICE (SOLID CARBON DIOXIDE), SEEN HERE BEING CONVERTED INTO A VAPOR, OFTEN IS USED TO CHANGE SUPERCOOLED WATER IN CLOUDS INTO ICE CRYSTALS IN THE WEATHER MODIFICATION TECHNIQUE CALLED CLOUD SEEDING. *(© M. Meadow/Photo Researchers. Reproduced by permission.)*

than their counterparts in developing countries and can evacuate hurricane regions more easily. Additionally, a wealthier and more powerful government is able to administer relief more quickly and in greater quantities than the leadership of small, poor nations.

Even so, hurricanes and typhoons have a vast impact wherever they strike, and perhaps for this reason a certain mystery and lore have developed around them. Reflective of this fascination is the practice of personalizing hurricanes. Originally, the U.S. National Weather Service gave them women's names, but in response to protests from feminist groups, by the late 1970s it had adopted a practice of using an alphabetic list of alternating male and female first names. The Weather Service draws up new lists each year to name the cyclones of the western Pacific and the Caribbean/Gulf regions.

FORECASTING AND CONTROLLING WEATHER

Until the twentieth century, people had little warning when a cyclone was coming. Early in that century, however, the establishment of hurricane-watch services offered the hope of early warning, and by the 1930s ships and weather balloons were used to provide readings on atmospheric conditions that might portend cyclones. In the 1940s airplanes and, later, radar further increased meteorologists' ability to monitor the atmosphere. Today a global network of weather satellites makes it possible to identify and track tropical cyclones from their earliest appearance as disturbances over the remote ocean.

From the time of the Greeks, at least, people have tried to forecast the weather. They attempted to do so with varying degrees of success, using means that included folklore, superstition, old wives' tales, traditional wisdom, instinct, intuition, and even a little bit of science. Sometimes this mixed bag yielded valuable information, such as the old and essentially accurate saying "Red sky at morning, sailors take warning; red sky at night, sailor's delight." Yet without true scientific methods of forecasting, would-be meteorologists were just shooting in the dark—as the sometimes inaccurate results of today's much more scientific forecasting shows.

A turning point came in the twentieth century, with the development of such monitoring systems as the cyclone-monitoring techniques we have mentioned. It says a great deal about just how young meteorology is as a science that one of its two founders as a modern discipline died only in the 1970s. This was Jakob Bjerknes (1897–1975), a Norwegian-born American meteorologist who, in the 1920s, established a network of weather stations with his father, the Norwegian physicist Vilhelm Bjerknes (1862–1951). Vilhelm proposed the idea of air masses, a pivotal concept in meteorology, while Jakob discovered the origin of cyclones.

MODERN WEATHER FORECASTING. Weather forecasting in the United States is the responsibility of the National Weather Service (NWS), which is part of the National Oceanic and Atmospheric Administration (NOAA) within the Department of Commerce. The NWS maintains a vast network of field offices and weather stations as well as nine National Centers for Environmental Prediction, each of which is focused on specific weather-related responsibilities. The complexity of the NWS organizational system hints at the greater

KEY TERMS

ATMOSPHERE: In general, an atmosphere is a blanket of gases surrounding a planet. Unless otherwise identified, however, the term refers to the atmosphere of Earth, which consists of nitrogen (78%), oxygen (21%), argon (0.93%), and other substances that include water vapor, carbon dioxide, ozone, and noble gases such as neon (0.07%).

CLIMATE: The pattern of weather conditions in a particular region over an extended period. Compare with *weather.*

CONDENSATION: The formation of a liquid from a vapor, usually as a result of a reduction in temperature.

CONVECTION: Vertical circulation that results from differences in density ultimately brought about by differences in temperature. Convection involves the transfer of heat through the motion of hot fluid (e.g., air) from one place to another.

CONVECTION CURRENT: The flow of material heated by means of convection.

EVAPORATION: The process whereby liquid water is converted into a gaseous state and transported to the atmosphere. When discussing the atmosphere and precipitation, evaporation generally is distinguished from transpiration. In this context, evaporation refers solely to the transfer of water from nonliving sources, such as the soil or the surface of a lake.

EVAPOTRANSPIRATION: The loss of water to the atmosphere via the combined (and related) processes of evaporation and transpiration.

HUMIDITY: The amount of water vapor in the air.

METEOROLOGY: The study of the atmosphere, weather, and weather prediction.

PRECIPITATION: When discussing the hydrologic cycle or meteorology, precipitation refers to the water, in liquid or solid form, that falls to the ground when the atmosphere has become saturated with moisture.

TRANSPIRATION: The process whereby moisture is transferred from living organisms to the atmosphere. A major portion of environmental moisture for precipitation comes from plants, which lose water through their stomata, small openings on the undersides of leaves.

UNSTABLE ATMOSPHERE: A term describing a situation in which air temperature drops rapidly with altitude. As a result, a packet of air continues to move upward owing to the temperature difference between it and the surrounding air.

WEATHER: The condition of the atmosphere at a given time and place. Compare with *climate.*

complexity of weather forecasting itself, which we touch on here in only the most cursory fashion.

A number of useful techniques and forms of technology are available to the modern meteorologist. Perhaps the best known of these is Doppler radar, used to track the movement of storm systems. By detecting the direction and velocity of raindrops or hail, Doppler radar can help the meteorologist determine the direction of winds, and thus predict weather patterns that will follow in the next minutes or hours. But Doppler radar can do more than simply detect a storm in progress: Doppler technology also aids meteorologists by interpreting wind direction. Other forms of weather-forecasting technology include NEXRAD (Next Generation Radar) and GOES (Geostationary Operational Environmental Satellite).

SOME TYPES OF WEATHER FORECAST. The simplest (and usually least accurate) type of forecast is called a persistent forecast and starts with the assumption that existing patterns will continue into the future. The problem with this notion, of course, is that with the complexity of weather, patterns are always changing. More reliable is the so-called trend method, based on the relationship between the movement of air masses and the larger weather patterns. This is a type of prediction familiar to most of us from weather maps displayed by TV meteorologists.

Similar to the trend method is the analogue method, which uses analogies (hence the name) between current weather maps and similar maps from the past. If a weather map for today closely matches the map of patterns that prevailed on a particular day three years ago, it is possible to make some predictions about the weather now by referring to conditions that developed at that time.

Meteorology may employ statistical probability, and when it comes to long-range forecasting, the mathematics may become considerably more complex. This is illustrated by the fact that chaos theory, one of the most challenging branches of modern mathematics, was the brainchild not of a mathematician but of the American meteorologist Edward Lorenz (1917–). An outgrowth of Lorenz's studies in atmospheric patterns, chaos theory is applied in studying extremely complex systems whose behavior appears random. (On a pop-culture note, chaos theory was the specialization of Ian Malcolm, the mathematician played by Jeff Goldblum in the film, *Jurassic Park*.)

"MAKING" WEATHER. People at the turn of the nineteenth century would have been flabbergasted to know that it would be possible by the twenty-first century to provide a reasonably accurate 24-hour weather forecast. They would have been even more astounded at the idea of a five-day forecast, which is common (though with varying degrees of accuracy) on most local and national weather reports. And, no doubt, they would have been "blown away"—to use a popular weather metaphor—at the concept that modern technology makes it possible even to exert some control over the weather.

Weather modification includes techniques such as cloud-seeding, which originated in the 1940s. By the beginning of the twenty-first century, several other methods of weather modification existed, among them frost prevention, fog and cloud dispersal, hurricane modification, hail suppression, and lighting suppression.

Cloud seeding necessitates the conversion of supercooled water in clouds—that is, water whose temperature has dropped below the freezing point without it actually freezing—into ice crystals. Dry ice (solid carbon dioxide) and silver iodide are the substances most commonly employed for this purpose. The result is, or at least can be, snow clouds. On the other hand, cloud-seeding techniques can be used for fog dispersal, another form of weather modification useful, for instance, in the skies over airports.

WHERE TO LEARN MORE

Allaby, Michael. *DK Guide to the Weather.* New York: Dorling Kindersley, 2000.

———. *Tornadoes and Other Dramatic Weather Systems.* New York: Dorling Kindersley, 2001.

The Atmosphere (Web site). <http://liftoff.msfc.nasa.gov/academy/space/atmosphere.html>.

Folk Lore Weather Forecasting (Web site). <http://www.shoal.net.au/~seabreeze/weather.html>.

National Weather Service/National Oceanic and Atmospheric Administration (Web site). <http://www.nws.noaa.gov/>.

Reynolds, Ross, ed. *Cambridge Guide to the Weather.* New York: Cambridge University Press, 2000.

Silverstein, Alvin, Virginia B. Silverstein, and Laura Nunn. *Weather and Climate.* Brookfield, Conn.: Twenty-First Century Books, 1998.

Stein, Paul, ed. *The Macmillan Encyclopedia of Weather.* New York: Macmillan Reference USA, 2001.

The Weather Channel (Web site). <http://www.weather.com/>.

Weather Underground (Web site). <http://www.wunderground.com/>.

CLIMATE

CONCEPT

Many people think of weather and climate as concepts that are very nearly synonymous, but this is far from the truth. Whereas *weather* is a term referring to the atmospheric conditions for a particular place at a particular time, *climate* describes the overall weather pattern of a region over an extended period. A spot on the South Carolina coast, for instance, is likely to be humid and prone to hurricanes; another place, in the Rocky Mountains, might be dry and windy; while yet another locale, in Hawaii, most likely is inclined to extremely mild and equable or unvarying temperatures. In each case, what has been described is climate; for the weather in any of those particular places, one would need to check the weather report on a specific day. Despite the fact that climate is a concept that encompasses the weather in a region over an extended period, it is still possible for climate itself to vary.

HOW IT WORKS

The Atmosphere

All weather takes place in the atmosphere, the uppermost of the four earth systems. A blanket of gases created in large part by the expulsion of elements from the geosphere through volcanic eruptions, the atmosphere sustains the biosphere through its components of oxygen and carbon dioxide. But it also recirculates water from the hydrosphere, a key function in its role as a weather-generating system. (See Earth Systems for more about the interactions among geosphere, biosphere, hydrosphere, and atmosphere.)

COMPOSITION OF THE ATMOSPHERE. Though we think of the atmosphere as being composed primarily of oxygen, the vast majority of it (78%) is nitrogen. The latter is a highly unreactive gas, meaning that it does not tend to bond chemically with other elements. (Nitrogen itself is discussed in considerably greater detail within the essay Nitrogen Cycle.) Therefore, the nitrogen in our air is really just "filler," the result of volcanic eruptions billions of years ago: unlike carbon dioxide, which dissolved in the water of Earth's oceans, the nitrogen simply stuck around in the atmosphere.

Air is not a chemical compound but a mixture, of which nitrogen and oxygen together account for 99%. Another 0.93% is argon, which is a noble gas, meaning that it, too, is highly unreactive. The last 0.07% comes from trace gases (i.e., gases in very small quantities), including two that are extremely important to the operation of Earth: carbon dioxide and water vapor. The first of these is a key component in the biosphere and in biogeochemical cycles (see Carbon Cycle), while the second is of enormous importance in weather and climate.

The Troposphere and Air

It may be surprising to learn that water vapor is such a small portion of the atmosphere; even more shocking is the tiny fraction that this amount of water (equivalent to many billions of gallons) constitutes in proportion to Earth's entire water supply. Furthermore, all our weather, which plays such an integral role in our lives, is generated in a very small portion of the atmosphere.

FOR EVERY CLIMATE REGION ON EARTH, THERE ARE PARTICULAR TYPES OF PLANTS AND ANIMALS THAT HAVE ADAPTED TO THE PREVAILING CONDITIONS. ANIMALS THAT LIVE IN POLAR REGIONS, LIKE THIS BEAR, TYPICALLY HAVE VERY SMALL EARS, GIVING THEM LESS SURFACE AREA TO EXPOSE TO THE COLD; THEY ALSO TEND TO HIBERNATE AS A MEANS OF ENDURING THE WINTER. *(© John Shaw/Photo Researchers. Reproduced by permission.)*

This is the troposphere, the lowest layer, which extends to about 10 mi. (16 km) above the surface of Earth. The height of the troposphere is about one-fifth the combined height of the troposphere, stratosphere, and mesosphere, the three atmospheric layers that contain air. (Beyond these layers are the thermosphere, which eventually dissolves into the exosphere and the emptiness of space, which is characterized precisely by its lack of an atmosphere.)

A HIGH CONCENTRATION OF AIR. Small as the troposphere is within the larger atmosphere, however, it contains 80% of the atmosphere's mass. The reason is that at higher altitudes, Earth exerts less gravitational attraction on the gas molecules that make up air. Thus, if one were to travel up through the outer layers, the amount of air in the atmosphere would simply shrink to nothingness because at that height, Earth does not exert enough gravitational force to hold in place those ultralight particles.

By contrast, in the troposphere, the gravitational force—a function of the distance between two objects, in this case, a gas molecule and Earth—is extremely strong. This results in a high concentration of gas near the surface. (Note the term *gas*: remember that air is a mixture, not a compound, so there is no such thing as an "air molecule." Instead, there are only molecules of oxygen [O_2], ozone [O_3], nitrogen [N_2], carbon dioxide, and water, as well as atoms of argon and other noble gases.)

In any case, the high concentration of gas creates pressure, which, along with several other factors, causes air to move. The result is weather, and weather patterns over a long enough period of time yield what we call climate. (See Weather for more about the components that make up weather as well as their interactions.)

THE ATMOSPHERIC SCIENCES

Weather is like a daily newspaper, whereas climate is like a book of history: whereas weather reflects the atmospheric conditions for a particular place at a particular time, climate is the overall pattern of weather over an extended period. This difference is reflected in the organization of the atmospheric sciences, the area of earth sciences concerned with atmospheric phenomena.

Principal among the atmospheric sciences are meteorology, which is the study of weather, and climatology, or the study of climate. Meteorology focuses on daily or even hourly changes in conditions within the troposphere and the lower stratosphere, the region just above it. By contrast, climatology is concerned with analyzing the weather in a particular area over periods as short as a month or as long as several million years.

Whereas meteorology involves daily weather forecasts and reports, efforts in climatology are directed toward explaining differences in climate around the earth. Climatologists also investigate how these differences are related to other aspects of the natural environment. Paleoclimatology is a specialized branch of climatology devoted to the study of climatic conditions in the distant past. Such study may require investigation of fossils and other materials that provide physical or chemical clues regarding the past climate and changes it experienced.

REAL-LIFE APPLICATIONS

Climates Zones and Organisms

One of the few ancient scientific thinkers whose work is still held in high regard was the Roman geographer Pomponius Mela (*fl. ca.* A.D. 44). In *De situ orbis,* Mela introduced the idea of five climate zones on Earth: northern frigid or cold, northern temperate or mild, torrid (very hot), southern temperate, and southern frigid. In a world that knew of no lands further north than Scandinavia (a semi-mythical place the Greeks had called "Thule") or further south than the lower Nile, Mela's designation of climate zones was extraordinarily accurate.

Only in the 1910s did the Russian climatologist Wladimir Köppen (1846–1940) improve on Mela's system, developing his own five-part climate designation. By Köppen's time, it was clear that there was not necessarily any climatic difference between northern and southern seasons, except inasmuch as the seasons in the Southern Hemisphere took place at opposite times of the year from those in the Northern Hemisphere. Therefore, he dispensed with all references to latitude except insofar as they related to position relative to the equator.

LARGE CLIMATE REGIONS. The Köppen system recognizes the zones of humid tropical, dry, humid mid-latitude with mild winters, humid mid-latitude with cold winters, and polar. Each of these larger categories is then subdivided into smaller climate types: for example, among the dry-climate group, there is a distinction between deserts and steppes, which are arid but not completely barren plains of a type found in Russia and central Asia.

Two factors, the average annual temperature and amount of precipitation, serve to differentiate climate types. Humid tropical, dry, and polar zones are fairly self-explanatory; as for the two humid mid-latitude types, they are distinguished by their distance from the equator. For instance, the southern United States would be an example of a humid mid-latitude region with mild winters, while the northeastern United States—New York and New England—would be considered a humid mid-latitude region with cold winters.

For every climate region on Earth, there are particular types of plants and animals that have adapted to the prevailing conditions. Animals that live in desert regions, for example, might have large ears to add a greater surface area for perspiration. On the other hand, animals in polar regions typically have very small ears, giving them less surface area to expose to the cold. Both camels and cacti are organisms well adapted to a desert climate: the camel can store large amounts of water and food in its hump, while the cactus requires little moisture to survive. Many a desert animal is nocturnal, allowing it to survive the heat, while most polar animals tend to hibernate as a means of enduring winter.

Microclimates

Not all climates necessarily spread across a whole desert or mountain range—or, in the case of Antarctica, with its decidedly polar climate, a whole continent. A very specific area either on Earth's surface or just a few feet or meters above or below it (i.e., up in the trees or in the soil) likewise can have its own microclimate. The very existence of a microclimate, with all its complexity, serves to show just how complex the larger Earth system is.

A particular microclimate, such as that of a forest or even a particular spot within a forest, has its own specific weather conditions: temperature, humidity, wind patterns, dew, frost, and evaporation or transpiration. Soil is a major factor influencing the quality of the microclimate: if the soil

BENI ABBES DUNES, SAHARA DESERT. SOIL IS A MAJOR FACTOR INFLUENCING THE QUALITY OF A MICROCLIMATE: SOIL THAT IS SANDY IN TEXTURE AND LIGHT IN COLOR IS LIKELY TO REFLECT MORE LIGHT AND HEAT. *(© V. Engelbert/Photo Researchers. Reproduced by permission.)*

is sandy in texture, it is likely to reflect more light and heat. The same is true if it is light in color. Also important is topography, an example being the rain shadow created by mountains. (See Mountains for an explanation of rain shadows.)

MICROCLIMATES IN ACTION. A beautiful example of microclimate in action can be found when hiking on Mount Santo Tomas outside Baguio in the northern Philippines. Nestled in the mountains, with still higher regions such as Santo Tomas nearby, Baguio has long been a popular resort owing to its cool temperatures. It is one of the few places in the Philippines where one can find evergreen conifers such as pine trees, and the semi-alpine climate makes Baguio a welcome relief from the heat of Manila and other regions further south.

Even though it lies in the northern portion of Luzon, the northernmost of the major islands in the Philippine archipelago, it is not latitude that gives Baguio its cool climate; instead, it is altitude. The same is true of other places around the world: Quito, Ecuador, for instance, which has an extremely cool climate because it is high in the Andes, even though it lies on the equator. (The 2001 movie *Proof of Life,* which portrays climatic conditions ranging from mild to cold, was filmed in and around Quito.)

On Santo Tomas, however, where the altitude is even greater than that of Baguio, it is possible to experience both the heat that characterizes most of the Philippines *and* the cool conditions of the mountains in northern Luzon. Standing at a particular spot along the mountainside, one can feel the heat that falls directly on the mountain, owing to its tropical latitude—which means that the sunlight reaches the surface at a more or less perpendicular angle. At the same time, one can feel the cool breeze that blows up the mountain as the result of its altitude. It is possible, in fact, to stand in such a way that one's back, to the mountain itself, is hot and sweaty while one's face enjoys the cool, moist highland breeze.

CAN CLIMATE CHANGE?

Though we have established that climate is a long-term weather pattern, that does not mean that climate itself is fixed and unchanging. Any number of factors can affect it. For example, the "fog" in London was once such an established fact that it became associated with that city in the way that wind is with Chicago.

But just as Chicago's status as the "Windy City" is something of a myth (in fact, there are more than a dozen cities more windy, though Chicago does experience powerful winds as a result of its proximity to Lake Michigan), the London fog was not actually what it seemed to be. The fog was not really fog at all but pollution produced by the burning of coal for heat. As coal gave way to gas and other types of heat during the twentieth century, the fog in London's climate changed.

HEATING UP OR COOLING DOWN? A less advantageous condition in the global climate may be a result of other technological changes, according to some scientists. The burning of fossil fuels, such as coal, natural gas, and petroleum, during the past century or more has put large amounts of carbon dioxide into the atmosphere. That part of the situation is fairly cut and dried; as to the potential result, the scientific community is divided.

Some scientists believe that higher concentrations of carbon dioxide will lead to a greater retention of heat in the atmosphere, resulting in a higher annual average global temperature. This phenomenon, known as global warming, has attracted considerable media attention owing to its promotion by environmental activists, including celebrities who embrace environmental causes. Yet the global warming position is far from the only interpretation that the facts suggest.

Other experts contend that high levels of carbon dioxide in the atmosphere will have the opposite effect. According to this view, the concentration of carbon dioxide may heat the atmosphere in the short term, but this will result only in a larger amount of evaporation from the oceans. This evaporation, in turn, will result in the formation of much larger cloud masses, which would have the effect of reflecting sunlight back into space—thus, in fact, *lowering* Earth's temperature.

The global warming position, with its media and celebrity support, has held the lead since the 1980s. For this reason, it is easy to forget that in the mid- to late 1970s, many environmentalists held an exactly opposite position. At that time, a series of extremely cold winters led to the claim that Earth was *cooling* and that in a few more centuries, a new ice age would ensue. In fact, this is a considerably more plausible position, given the fact that Earth has experienced countless ice ages in the past.

BLOCKING OUT THE SUN. Of course, it could be maintained (as advocates of global warming do) that the present situation of massive fossil-fuel burning has never occurred in Earth's history. This is certainly true, but there is less credibility in the claim that never before has so much carbon dioxide been introduced to the atmosphere. In fact, billions of years ago, volcanoes belched vast quantities of the gas into the region above Earth's surface, but because carbon dioxide is highly water-soluble, most of it dissolved into the oceans' waters.

This points to one of the means by which a rapid climate change, in particular, a sudden cooling, is brought about: by a volcanic eruption or other phenomenon that places enormous amounts of dust and ash in the air. In so doing, it reduces the amount of solar radiation that reaches Earth's surface, causing temperatures to drop. An extreme example of this may have occurred about 65 million years ago, as the result of a meteorite hitting Earth and causing rapid climatic changes that brought about the mass extinction of the dinosaurs and other life-forms (see Paleontology).

A more benign example of atmospheric blockage occurred in 1991, after Mount Pinatubo in the Philippines erupted. The eruption appar-

NEW YORK CITY HAS A MASSIVE CONCENTRATION OF HUMANS, MACHINES, AND CONCRETE IN A VERY SMALL AREA. CITIES WITHOUT ADEQUATE GREEN SPACE, SUCH AS NEW YORK'S CENTRAL PARK (SHOWN HERE), BECOME GIANT REFLECTORS, AND THE PRESENCE OF SMOG AND HEAT FROM CARS AND OTHER MACHINES ADDS TO THE UNHEALTHY ENVIRONMENT. *(© Rafael Macia/Photo Researchers. Reproduced by permission.)*

ently released so much ash and other material into the atmosphere that it temporarily reversed a general warming trend that had prevailed during the 1980s. By the mid-1990s, however, the materials from Mount Pinatubo appeared to have settled out of the atmosphere, thus causing a return to early climate trends.

HEATING UP A MICROCLIMATE. Despite the questionable nature of some environmentalists' claims concerning the global climate, environmentalists are absolutely correct in noting the effects of human civilization, development, and technology on microclimates. An excellent example is New York City. Though it is located in the humid temperate climate zone described earlier, with cold winters, the city has less snowfall—and hotter summers—than less populated areas of New York State or even areas in Pennsylvania that lie on the same latitude.

The difference, of course, lies in the fact that New York City is one of the planet's greatest municipalities, a massive concentration of humans, machines, and concrete in a very small area. Heat-sensitive satellite imaging equipment routinely produces images of the United States that show great areas of heat around the major cities, particularly New York. Concrete, buildings, roads, and the like constitute what developers call impervious surface, meaning that rain is not supposed to leach through these surfaces to the soil; rather, it will remain on the surface as runoff. But sunlight cannot gather in puddles or run off into storm drains: it can only reflect, and thus a big city is a huge mirror to the Sun's rays.

A major city, from the standpoint of climate, is rather like a person who places a mirrored piece of metal around his or her face to "soak up the Sun's rays"—an extremely unhealthy, dangerous, and inadvisable practice. Cities without adequate green space, such as parks, become giant reflectors, and the presence of smog and heat from cars and other machines only adds to the unhealthy, artificially heated environment.

It is particularly unfortunate to see such changes in a city such as Atlanta, which until about 1980 was a fairly sleepy town noted for its abundance of trees. Atlanta has long since become a boomtown, and the trees of its metropolitan area are being massacred at an alarming rate. Not only is this denuding the city and suburbs (and in the process, some would say, stripping its last vestiges of charm), but it is raising Atlanta's average temperature—which was already high as a result of the city's latitude.

EARTH, SPACE, AND ICE AGES

If Earth were a horse on a carousel, with the Sun at the spinning center of the merry-go-round, the horse representing our planet would not be sitting upright, at a 90° angle to the carousel floor. Rather, it would be tilted off the perpendicular angle by 23.5°, a fact responsible for many features of the global climate and microclimates of Earth.

At certain times of the year, rays from the Sun strike either the Northern or Southern hemisphere more directly than at other times. Summer occurs in the Northern Hemisphere during the middle of the year, because, at that point in the planet's movement around the Sun, the rays of the Sun strike the Northern Hemisphere at an

KEY TERMS

ATMOSPHERE: In general, an atmosphere is a blanket of gases surrounding a planet. Unless otherwise identified, however, the term refers to the atmosphere of Earth, which consists of nitrogen (78%), oxygen (21%), argon (0.93%), and other substances that include water vapor, carbon dioxide, ozone, and noble gases such as neon (0.07%).

ATMOSPHERIC SCIENCES: A major division of the earth sciences, distinguished from geoscience and the hydrologic sciences by its concentration on atmospheric phenomena. Among the atmospheric sciences are meteorology and climatology.

CLIMATE: The pattern of weather conditions in a particular region over an extended period. Compare with *weather.*

CLIMATOLOGY: An area of the atmospheric sciences devoted to studying the weather in a particular region or regions over periods as short as a month or as long as several million years.

METEOROLOGY: The study of the atmosphere, weather, and weather prediction.

MICROCLIMATE: The climate of a very specific region a few feet or meters above or below Earth's surface. The size of microclimates is undefined and variable, but they are most definitely smaller than the regions (a state, a country, even a continent) over which a particular climate is said to prevail.

TROPOSPHERE: The atmospheric layer closest to the surface, extending upward approximately 10 mi. (16 km). The troposphere contains about 80% of the air in the atmosphere and is the region where weather occurs.

WEATHER: The condition of the atmosphere at a given time and place. Compare with *climate.*

angle close to the perpendicular. The same happens in the Southern Hemisphere around the end of the year, when that section of the planet receives solar radiation at a nearly perpendicular angle.

Another important factor that relates Earth's position in space to its climate and microclimates is its pattern of movement around the Sun. The carousel analogy is a flawed one, because a carousel is round; Earth's orbit, in fact, describes an ellipse, or oval. The reason is that the Sun exerts a greater gravitational pull on Earth at different parts of its orbital path, and the result is that at its closest approach to the Sun, Earth receives more solar energy than it does when it is farthest away from it. This has little do with seasons: the closest approach occurs in January, winter in the Northern Hemisphere. Conversely, in July, the hottest time of year for the Northern Hemisphere, Earth is at its furthest point from the Sun.

EXPLAINING ICE AGES. It is possible that Earth's position relative to the Sun may explain those dramatic periods of cooling known as *ice ages,* when much of Earth's water freezes, larger amounts of land are exposed, and some life-forms die off while new ones develop. There have been numerous ice ages in Earth's history, and what we call *the* Ice Age, which ended about 10,000 years ago, was just the last ice age.

That one was particularly significant, of course, not only because it was most recent, occurring as it did on the eve of civilization's beginnings, but because it was a formative juncture in human development. During those thousands of years, human hunter-gatherer society and Paleolithic technology developed to its highest point. Also, many of the significant migrations of people took place, most notably the movement of Siberian tribes eastward across the land bridge of what is now the Bering Strait.

What causes ice ages and other large-scale climate changes? In the 1930s the Serbian astrophysicist Milutin Milankovitch (1879–1958) put forward a theory maintaining that changes in the pattern of Earth's orbit around the sun, as well as in Earth's inclination to the plane of its orbit, could explain these changes in climate. The planet's axis of inclination (that is, its angle in relation to the plane of orbit) changes over a period of about 22,000 years, a cycle known as the *precession of the equinox.* As a result, first one hemisphere and then the other is be pointed toward the Sun.

Milankovitch evolved a complex theory that took into account precession of the equinoxes, changes in the shape of Earth's orbit, and changes in its orbital tilt. The theory has been improved and modified numerous times since the 1930s, but, in general, Milankovitch's ideas still provide climatologists with a basic understanding as to how large changes in Earth's climate take place.

WHERE TO LEARN MORE

The Atmosphere (Web site). <http://liftoff.msfc.nasa.gov/academy/space/atmosphere.html>.

Climate Prediction Center—National Oceanic and Atmospheric Administration/NOAA (Web site). <http://www.cpc.ncep.noaa.gov/>.

Elsner, James B., and A. Birol Kara. *Hurricanes of the North Atlantic: Climate and Society.* New York: Oxford University Press, 1999.

National Climatic Data Center—National Oceanic and Atmospheric Administration/NOAA (Web site). <http://lwf.ncdc.noaa.gov/oa/ncdc.html>.

Philander, S. George. *Is the Temperature Rising?: The Uncertain Science of Global Warming.* Princeton, NJ: Princeton University Press, 1998.

Resnick, Abraham. *Due to the Weather: Ways the Elements Affect Our Lives.* Westport, CT: Greenwood Press, 2000.

Silverstein, Alvin, Virginia B. Silverstein, and Laura Nunn. *Weather and Climate.* Brookfield, CT: Twenty-First Century Books, 1998.

Stevens, William K. *The Change in the Weather: People, Weather, and the Science of Climate.* New York: Delacorte Press, 1999.

United Nations Framework Convention on Climate Change/UNFCCC (Web site). <http://www.unfccc.de/>.

World Climate: Weather Rainfall and Temperature Data (Web site). <http://www.worldclimate.com/>.

INDEX OF EVERYDAY THINGS

The "everyday" items discussed in this series (volumes 1-4) have been categorized into 26 general subject areas, and are listed below. Boldface type indicates main entry volume and page numbers. Italic type indicates photo and illustration volume and page numbers.

AGRICULTURE

ANIMALS

ART

AUTOMOBILES

AVIATION

CIVIL ENGINEERING

COMMUNICATIONS

COMPUTERS

CRIME AND POLICE WORK

ELECTRONICS

ENERGY

ENVIRONMENT

FOOD AND DRINK

GEOLOGY

HOUSEHOLD PRODUCTS

MACHINES

MANUFACTURING

MEDICINE

MILITARY

MUSIC

NAVIGATION

SPORTS AND HOBBIES

TEXTILES

TRANSPORTATION

WEAPONS

WEATHER

CUMULATIVE GENERAL SUBJECT INDEX

This index contains items from volumes 1-4 of the series.
Boldface type indicates main entry volume and page numbers.
Italic type indicates photo and illustration volume and page numbers.

A

B

C

D

E

F

G

I

M

N

P

Q

R

S

T

V

X

Y

Z

SCIENCE OF EVERYDAY THINGS

VOLUME 2 REAL-LIFE PHYSICS

A Schlager Information Group Book
Neil Schlager, Editor
Written by Judson Knight

Gale Group Staff
Kimberley A. McGrath, *Senior Editor*
Maria Franklin, *Permissions Manager*
Margaret A. Chamberlain, *Permissions Specialist*
Shalice Shah-Caldwell, *Permissions Associate*
Mary Beth Trimper, *Manager, Composition and Electronic Prepress*
Evi Seoud, *Assistant Manager, Composition and Electronic Prepress*
Dorothy Maki, *Manufacturing Manager*
Rita Wimberley, *Buyer*
Michelle DiMercurio, *Senior Art Director*
Barbara J. Yarrow, *Manager, Imaging and Multimedia Content*
Robyn V. Young, *Project Manager, Imaging and Multimedia Content*
Leitha Etheridge-Sims, Mary K. Grimes, and David G. Oblender, *Image Catalogers*
Pam A. Reed, *Imaging Coordinator*
Randy Bassett, *Imaging Supervisor*
Robert Duncan, *Senior Imaging Specialist*
Dan Newell, *Imaging Specialist*

ISBN 0-7876-5631-3 (set)
0-7876-5632-1 (vol. 1) 0-7876-5634-8 (vol. 3)
0-7876-5633-X (vol. 2) 0-7876-5635-6 (vol. 4)

Printed in the United States of America
10 9 8 7 6 5 4 3 2 1

ITP™

Library of Congress Cataloging-in-Publication Data

Knight, Judson.
Science of everyday things / written by Judson Knight, Neil Schlager, editor.
p. cm.
Includes bibliographical references and indexes.
Contents: v. 1. Real-life chemistry – v. 2 Real-life physics.
ISBN 0-7876-5631-3 (set : hardcover) – ISBN 0-7876-5632-1 (v. 1) – ISBN 0-7876-5633-X (v. 2)
1. Science–Popular works. I. Schlager, Neil, 1966-II. Title.

Q162.K678 2001
500–dc21 2001050121

CONTENTS

INTRODUCTION

Overview of the Series

Welcome to *Science of Everyday Things*. Our aim is to explain how scientific phenomena can be understood by observing common, real-world events. From luminescence to echolocation to buoyancy, the series will illustrate the chief principles that underlay these phenomena and explore their application in everyday life. To encourage cross-disciplinary study, the entries will draw on applications from a wide variety of fields and endeavors.

Science of Everyday Things initially comprises four volumes:

Volume 1: *Real-Life Chemistry*
Volume 2: *Real-Life Physics*
Volume 3: *Real-Life Biology*
Volume 4: *Real-Life Earth Science*

Future supplements to the series will expand coverage of these four areas and explore new areas, such as mathematics.

Arrangement of Real Life Physics

This volume contains 40 entries, each covering a different scientific phenomenon or principle. The entries are grouped together under common categories, with the categories arranged, in general, from the most basic to the most complex. Readers searching for a specific topic should consult the table of contents or the general subject index.

Within each entry, readers will find the following rubrics:

- **Concept** Defines the scientific principle or theory around which the entry is focused.
- **How It Works** Explains the principle or theory in straightforward, step-by-step language.
- **Real-Life Applications** Describes how the phenomenon can be seen in everyday events.
- **Where to Learn More** Includes books, articles, and Internet sites that contain further information about the topic.

Each entry also includes a "Key Terms" section that defines important concepts discussed in the text. Finally, each volume includes numerous illustrations, graphs, tables, and photographs.

In addition, readers will find the comprehensive general subject index valuable in accessing the data.

About the Editor, Author, and Advisory Board

Neil Schlager and Judson Knight would like to thank the members of the advisory board for their assistance with this volume. The advisors were instrumental in defining the list of topics, and reviewed each entry in the volume for scientific accuracy and reading level. The advisors include university-level academics as well as high school teachers; their names and affiliations are listed elsewhere in the volume.

NEIL SCHLAGER is the president of Schlager Information Group Inc., an editorial services company. Among his publications are *When Technology Fails* (Gale, 1994); *How Products Are Made* (Gale, 1994); the *St. James Press Gay and Lesbian Almanac* (St. James Press, 1998); *Best Literature By and About Blacks* (Gale, 2000); *Contemporary Novelists, 7th ed.* (St. James Press, 2000); and *Science and Its Times* (7 vols., Gale, 2000-2001). His publications have won numerous awards, including three RUSA awards from the American Library Association, two Reference Books Bulletin/Booklist Editors' Choice awards, two New York Public Library Outstanding Reference awards, and a CHOICE award for best academic book.

Judson Knight is a freelance writer, and author of numerous books on subjects ranging from science to history to music. His work on science titles includes *Science, Technology, and Society, 2000 B.C.-A.D. 1799* (U*X*L, 2002), as well as extensive contributions to Gale's seven-volume *Science and Its Times* (2000-2001). As a writer on history, Knight has published *Middle Ages Reference Library* (2000), *Ancient Civilizations* (1999), and a volume in U*X*L's *African American Biography* series (1998). Knight's publications in the realm of music include *Parents Aren't Supposed to Like It* (2001), an overview of contemporary performers and genres, as well as *Abbey Road to Zapple Records: A Beatles Encyclopedia* (Taylor, 1999). His wife, Deidre Knight, is a literary agent and president of the Knight Agency. They live in Atlanta with their daughter Tyler, born in November 1998.

COMMENTS AND SUGGESTIONS

Your comments on this series and suggestions for future editions are welcome. Please write: The Editor, *Science of Everyday Things,* Gale Group, 27500 Drake Road, Farmington Hills, MI 48331.

ADVISORY BOARD

GENERAL CONCEPTS

FRAME OF REFERENCE

CONCEPT

Among the many specific concepts the student of physics must learn, perhaps none is so deceptively simple as frame of reference. On the surface, it seems obvious that in order to make observations, one must do so from a certain point in space and time. Yet, when the implications of this idea are explored, the fuller complexities begin to reveal themselves. Hence the topic occurs at least twice in most physics textbooks: early on, when the simplest principles are explained—and near the end, at the frontiers of the most intellectually challenging discoveries in science.

HOW IT WORKS

There is an old story from India that aptly illustrates how frame of reference affects an understanding of physical properties, and indeed of the larger setting in which those properties are manifested. It is said that six blind men were presented with an elephant, a creature of which they had no previous knowledge, and each explained what he thought the elephant was.

The first felt of the elephant's side, and told the others that the elephant was like a wall. The second, however, grabbed the elephant's trunk, and concluded that an elephant was like a snake. The third blind man touched the smooth surface of its tusk, and was impressed to discover that the elephant was a hard, spear-like creature. Fourth came a man who touched the elephant's legs, and therefore decided that it was like a tree trunk. However, the fifth man, after feeling of its tail, disdainfully announced that the elephant was nothing but a frayed piece of rope. Last of all, the sixth blind man, standing beside the elephant's slowly flapping ear, felt of the ear itself and determined that the elephant was a sort of living fan.

These six blind men went back to their city, and each acquired followers after the manner of religious teachers. Their devotees would then argue with one another, the snake school of thought competing with adherents of the fan doctrine, the rope philosophy in conflict with the tree trunk faction, and so on. The only person who did not join in these debates was a seventh blind man, much older than the others, who had visited the elephant after the other six.

While the others rushed off with their separate conclusions, the seventh blind man had taken the time to pet the elephant, to walk all around it, to smell it, to feed it, and to listen to the sounds it made. When he returned to the city and found the populace in a state of uproar between the six factions, the old man laughed to himself: he was the only person in the city who was not convinced he knew exactly what an elephant was like.

Understanding Frame of Reference

The story of the blind men and the elephant, within the framework of Indian philosophy and spiritual beliefs, illustrates the principle of syadvada. This is a concept in the Jain religion related to the Sanskrit word *syat*, which means "may be." According to the doctrine of syadvada, no judgment is universal; it is merely a function of the circumstances in which the judgment is made.

On a complex level, syadvada is an illustration of relativity, a topic that will be discussed later; more immediately, however, both syadvada and the story of the blind men beautifully illustrate the ways that frame of reference affects perceptions. These are concerns of fundamental importance both in physics and philosophy, disciplines that once were closely allied until each became more fully defined and developed. Even in the modern era, long after the split between the two, each in its own way has been concerned with the relationship between subject and object.

These two terms, of course, have numerous definitions. Throughout this book, for instance, the word "object" is used in a very basic sense, meaning simply "a physical object" or "a thing." Here, however, an object may be defined as something that is perceived or observed. As soon as that definition is made, however, a flaw becomes apparent: nothing is just perceived or observed in and of itself—there has to be someone or something that actually perceives or observes. That something or someone is the subject, and the perspective from which the subject perceives or observes the object is the subject's frame of reference.

AMERICA AND CHINA: FRAME OF REFERENCE IN PRACTICE. An old joke—though not as old as the story of the blind men—goes something like this: "I'm glad I wasn't born in China, because I don't speak Chinese." Obviously, the humor revolves around the fact that if the speaker were born in China, then he or she would have grown up speaking Chinese, and English would be the foreign language.

The difference between being born in America and speaking English on the one hand—even if one is of Chinese descent—or of being born in China and speaking Chinese on the other, is not just a contrast of countries or languages. Rather, it is a difference of worlds—a difference, that is, in frame of reference.

Indeed, most people would see a huge distinction between an English-speaking American and a Chinese-speaking Chinese. Yet to a visitor from another planet—someone whose frame of reference would be, quite literally, otherworldly—the American and Chinese would have much more in common with each other than either would with the visitor.

The View from Outside and Inside

Now imagine that the visitor from outer space (a handy example of someone with no preconceived ideas) were to land in the United States. If the visitor landed in New York City, Chicago, or Los Angeles, he or she would conclude that America is a very crowded, fast-paced country in which a number of ethnic groups live in close proximity. But if the visitor first arrived in Iowa or Nebraska, he or she might well decide that the United States is a sparsely populated land, economically dependent on agriculture and composed almost entirely of Caucasians.

A landing in San Francisco would create a falsely inflated impression regarding the number of Asian Americans or Americans of Pacific Island descent, who actually make up only a small portion of the national population. The same would be true if one first arrived in Arizona or New Mexico, where the Native American population is much higher than for the nation as a whole. There are numerous other examples to be made in the same vein, all relating to the visitors' impressions of the population, economy, climate, physical features, and other aspects of a specific place. Without consulting some outside reference point—say, an almanac or an atlas—it would be impossible to get an accurate picture of the entire country.

The principle is the same as that in the story of the blind men, but with an important distinction: an elephant is an example of an identifiable species, whereas the United States is a unique entity, not representative of some larger class of thing. (Perhaps the only nation remotely comparable is Brazil, also a vast land settled by outsiders and later populated by a number of groups.) Another important distinction between the blind men story and the United States example is the fact that the blind men were viewing the elephant from outside, whereas the visitor to America views it from inside. This in turn reflects a difference in frame of reference relevant to the work of a scientist: often it is possible to view a process, event, or phenomenon from outside; but sometimes one must view it from inside—which is more challenging.

Frame of Reference in Science

Philosophy (literally, "love of knowledge") is the most fundamental of all disciplines: hence, most persons who complete the work for a doctorate receive a "doctor of philosophy" (Ph.D.) degree. Among the sciences, physics—a direct offspring of philosophy, as noted earlier—is the most fundamental, and frame of reference is among its most basic concepts.

Hence, it is necessary to take a seemingly backward approach in explaining how frame of reference works, examining first the broad applications of the principle and then drawing upon its specific relation to physics. It makes little sense to discuss first the ways that physicists apply frame of reference, and only then to explain the concept in terms of everyday life. It is more meaningful to relate frame of reference first to familiar, or at least easily comprehensible, experiences—as has been done.

At this point, however, it is appropriate to discuss how the concept is applied to the sciences. People use frame of reference every day—indeed, virtually every moment—of their lives, without thinking about it. Rare indeed is the person who "walks a mile in another person's shoes"—that is, someone who tries to see events from the viewpoint of another. Physicists, on the other hand, have to be acutely aware of their frame of reference. Moreover, they must "rise above" their frame of reference in the sense that they have to take it into account in making calculations. For physicists in particular, and scientists in general, frame of reference has abundant "real-life applications."

LINES OF LONGITUDE ON EARTH ARE MEASURED AGAINST THE LINE PICTURED HERE: THE "PRIME MERIDIAN" RUNNING THROUGH GREENWICH, ENGLAND. AN IMAGINARY LINE DRAWN THROUGH THAT SPOT MARKS THE Y-AXIS FOR ALL VERTICAL COORDINATES ON EARTH, WITH A VALUE OF 0° ALONG THE X-AXIS, WHICH IS THE EQUATOR. THE PRIME MERIDIAN, HOWEVER, IS AN ARBITRARY STANDARD THAT DEPENDS ON ONE'S FRAME OF REFERENCE. *(Photograph by Dennis di Cicco/Corbis. Reproduced by permission.)*

REAL-LIFE APPLICATIONS

POINTS AND GRAPHS

There is no such thing as an absolute frame of reference—that is, a frame of reference that is fixed, and not dependent on anything else. If the entire universe consisted of just two points, it would be impossible (and indeed irrelevant) to say which was to the right of the other. There would be no right and left: in order to have such a distinction, it is necessary to have a third point from which to evaluate the other two points.

As long as there are just two points, there is only one dimension. The addition of a third point—as long as it does not lie along a straight line drawn through the first two points—creates two dimensions, length and width. From the frame of reference of any one point, then, it is possible to say which of the other two points is to the right.

Clearly, the judgment of right or left is relative, since it changes from point to point. A more absolute judgment (but still not a completely absolute one) would only be possible from the frame of reference of a fourth point. But to constitute a new dimension, that fourth point could not lie on the same plane as the other three points—more specifically, it should not be possible to create a single plane that encompasses all four points.

Assuming that condition is met, however, it then becomes easier to judge right and left. Yet right and left are never fully absolute, a fact easily illustrated by substituting people for points. One may look at two objects and judge which is to the right of the other, but if one stands on one's head, then of course right and left become reversed.

Of course, when someone is upside-down, the correct orientation of left and right is still fairly obvious. In certain situations observed by physicists and other scientists, however, orientation is not so simple. It then becomes necessary to assign values to various points, and for this, scientists use tools such as the Cartesian coordinate system.

COORDINATES AND AXES. Though it is named after the French mathematician and philosopher René Descartes (1596-1650), who first described its principles, the Cartesian system owes at least as much to Pierre de Fermat (1601-1665). Fermat, a brilliant French amateur mathematician—amateur in the sense that he was not trained in mathematics, nor did he earn a living from that discipline—greatly developed the Cartesian system.

A coordinate is a number or set of numbers used to specify the location of a point on a line, on a surface such as a plane, or in space. In the Cartesian system, the x-axis is the horizontal line of reference, and the y-axis the vertical line of reference. Hence, the coordinate (0, 0) designates the point where the x- and y-axes meet. All numbers to the right of 0 on the x-axis, and above 0 on the y-axis, have a positive value, while those to the left of 0 on the x-axis, or below 0 on the y-axis have a negative value.

This version of the Cartesian system only accounts for two dimensions, however; therefore, a z-axis, which constitutes a line of reference for the third dimension, is necessary in three-dimensional graphs. The z-axis, too, meets the x- and y-axes at (0, 0), only now that point is designated as (0, 0, 0).

In the two-dimensional Cartesian system, the x-axis equates to "width" and the y-axis to "height." The introduction of a z-axis adds the dimension of "depth"—though in fact, length, width, and height are all relative to the observer's frame of reference. (Most representations of the three-axis system set the x- and y-axes along a horizontal plane, with the z-axis perpendicular to them.) Basic studies in physics, however, typically involve only the x- and y-axes, essential to plotting graphs, which, in turn, are integral to illustrating the behavior of physical processes.

THE TRIPLE POINT. For instance, there is a phenomenon known as the "triple point," which is difficult to comprehend unless one sees it on a graph. For a chemical compound such as water or carbon dioxide, there is a point at which it is simultaneously a liquid, a solid, and a vapor. This, of course, seems to go against common sense, yet a graph makes it clear how this is possible.

Using the x-axis to measure temperature and the y-axis pressure, a number of surprises become apparent. For instance, most people associate water as a vapor (that is, steam) with very high temperatures. Yet water can also be a vapor—for example, the mist on a winter morning—at relatively low temperatures and pressures, as the graph shows.

The graph also shows that the higher the temperature of water vapor, the higher the pressure will be. This is represented by a line that curves upward to the right. Note that it is not a straight line along a 45° angle: up to about 68°F (20°C), temperature increases at a somewhat greater rate than pressure does, but as temperature gets higher, pressure increases dramatically.

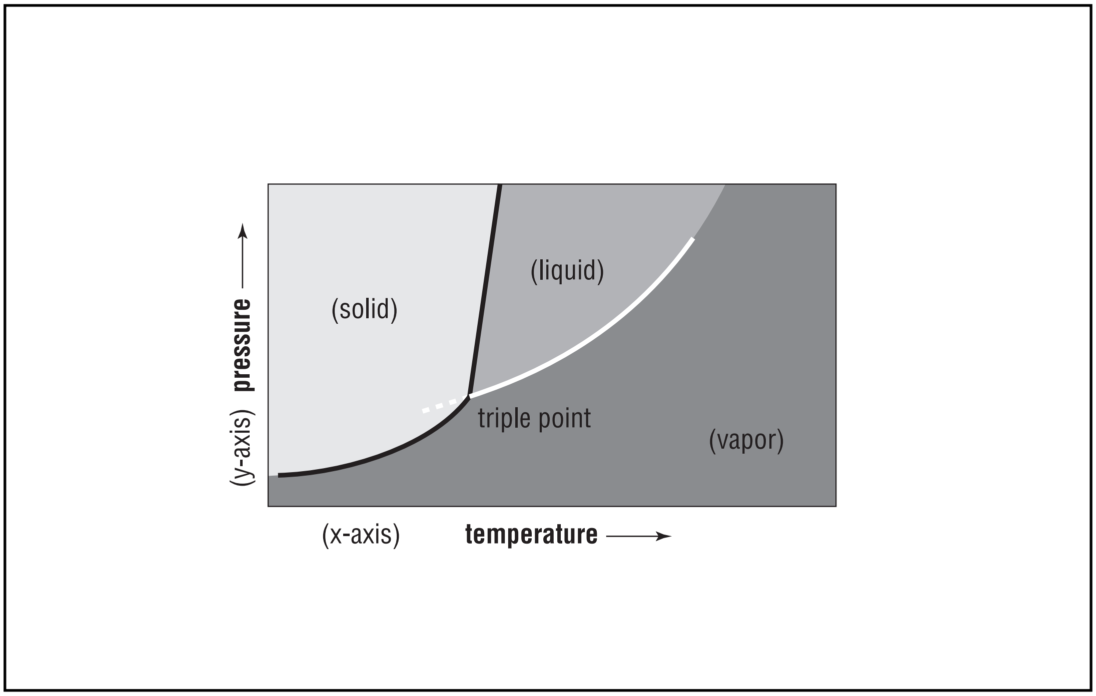
As everyone knows, at relatively low temperatures water is a solid—ice. Pressure, however, is relatively high: thus on a graph, the values of temperatures and pressure for ice lie above the vaporization curve, but do not extend to the right of 32°F (0°C) along the x-axis. To the right of 32°F, but above the vaporization curve, are the coordinates representing the temperature and pressure for water in its liquid state.

Water has a number of unusual properties, one of which is its response to high pressures and low temperatures. If enough pressure is applied, it is possible to melt ice—thus transforming it from a solid to a liquid—at temperatures below the normal freezing point of 32°F. Thus, the line that divides solid on the left from liquid on the right is not exactly parallel to the y-axis: it slopes gradually toward the y-axis, meaning that at ultra-high pressures, water remains liquid even though it is well below the freezing point.

Nonetheless, the line between solid and liquid has to intersect the vaporization curve somewhere, and it does—at a coordinate slightly above freezing, but well below normal atmospheric pressure. This is the triple point, and though "common sense" might dictate that a thing cannot possibly be solid, liquid, and vapor all at once, a graph illustrating the triple point makes it clear how this can happen.

Numbers

In the above discussion—and indeed throughout this book—the existence of the decimal, or base-



This Cartesian coordinate graph shows how a substance such as water could experience a triple point—a point at which it is simultaneously a liquid, a solid, and a vapor.

10, numeration system is taken for granted. Yet that system is a wonder unto itself, involving a complicated interplay of arbitrary and real values. Though the value of the number 10 is absolute, the expression of it (and its use with other numbers) is relative to a frame of reference: one could just as easily use a base-12 system.

Each numeration system has its own frame of reference, which is typically related to aspects of the human body. Thus throughout the course of history, some societies have developed a base-2 system based on the two hands or arms of a person. Others have used the fingers on one hand (base-5) as their reference point, or all the fingers and toes (base-20). The system in use throughout most of the world today takes as its frame of reference the ten fingers used for basic counting.

COEFFICIENTS. Numbers, of course, provide a means of assigning relative values to a variety of physical characteristics: length, mass, force, density, volume, electrical charge, and so on. In an expression such as "10 meters," the numeral 10 is a coefficient, a number that serves as a measure for some characteristic or property. A coefficient may also be a factor against which other values are multiplied to provide a desired result.

For instance, the figure 3.141592, better known as pi (π), is a well-known coefficient used in formulae for measuring the circumference or area of a circle. Important examples of coefficients in physics include those for static and sliding friction for any two given materials. A coefficient is simply a number—not a value, as would be the case if the coefficient were a measure of something.

Standards of Measurement

Numbers and coefficients provide a convenient lead-in to the subject of measurement, a practical example of frame of reference in all sciences—and indeed, in daily life. Measurement always requires a standard of comparison: something that is fixed, against which the value of other things can be compared. A standard may be arbitrary in its origins, but once it becomes fixed, it provides a frame of reference.

Lines of longitude, for instance, are measured against an arbitrary standard: the "Prime Meridian" running through Greenwich, England. An imaginary line drawn through that spot marks the line of reference for all longitudinal measures on Earth, with a value of 0°. There is nothing special about Greenwich in any profound scientific sense; rather, its place of importance reflects that of England itself, which ruled the seas and indeed much of the world at the time the Prime Meridian was established.

The Equator, on the other hand, has a firm scientific basis as the standard against which all lines of latitude are measured. Yet today, the coordinates of a spot on Earth's surface are given in relation to both the Equator and the Prime Meridian.

CALIBRATION. Calibration is the process of checking and correcting the performance of a measuring instrument or device against the accepted standard. America's preeminent standard for the exact time of day, for instance, is the United States Naval Observatory in Washington, D.C. Thanks to the Internet, people all over the country can easily check the exact time, and correct their clocks accordingly.

There are independent scientific laboratories responsible for the calibration of certain instruments ranging from clocks to torque wrenches, and from thermometers to laser beam power analyzers. In the United States, instruments or devices with high-precision applications—that is, those used in scientific studies, or by high-tech industries—are calibrated according to standards established by the National Institute of Standards and Technology (NIST).

THE VALUE OF STANDARDIZATION TO A SOCIETY. Standardization of weights and measures has always been an important function of government. When Ch'in Shih-huang-ti (259-210 B.C.) united China for the first time, becoming its first emperor, he set about standardizing units of measure as a means of providing greater unity to the country—thus making it easier to rule.

More than 2,000 years later, another empire—Russia—was negatively affected by its failure to adjust to the standards of technologically advanced nations. The time was the early twentieth century, when Western Europe was moving forward at a rapid pace of industrialization. Russia, by contrast, lagged behind—in part because its failure to adopt Western standards put it at a disadvantage.

Train travel between the West and Russia was highly problematic, because the width of railroad tracks in Russia was different than in Western Europe. Thus, adjustments had to be performed on trains making a border crossing, and this created difficulties for passenger travel. More importantly, it increased the cost of transporting freight from East to West.

Russia also used the old Julian calendar, as opposed to the Gregorian calendar adopted throughout much of Western Europe after 1582. Thus October 25, 1917, in the Julian calendar of old Russia translated to November 7, 1917 in the Gregorian calendar used in the West. That date was not chosen arbitrarily: it was then that Communists, led by V. I. Lenin, seized power in the weakened former Russian Empire.

METHODS OF DETERMINING STANDARDS. It is easy to understand, then, why governments want to standardize weights and measures—as the U.S. Congress did in 1901, when it established the Bureau of Standards (now NIST) as a nonregulatory agency within the Commerce Department. Today, NIST maintains a wide variety of standard definitions regarding mass, length, temperature, and so forth, against which other devices can be calibrated.

Note that NIST keeps on hand definitions rather than, say, a meter stick or other physical model. When the French government established the metric system in 1799, it calibrated the value of a kilogram according to what is now known as the International Prototype Kilogram, a platinum-iridium cylinder housed near Sèvres in France. In the years since then, the trend has moved away from such physical expressions of standards, and toward standards based on a constant figure. Hence, the meter is defined as the distance light travels in a vacuum (an area of space devoid of air or other matter) during the interval of 1/299,792,458 of a second.

METRIC VS. BRITISH. Scientists almost always use the metric system, not because it is necessarily any less arbitrary than the British or English system (pounds, feet, and so on), but because it is easier to use. So universal is the metric system within the scientific community that it is typically referred to simply as SI, an abbreviation of the French *Système International d'Unités*—that is, "International System of Units."

The British system lacks any clear frame of reference for organizing units: there are 12 inches in a foot, but 3 feet in a yard, and 1,760 yards in a mile. Water freezes at 32°F instead of 0°, as it does in the Celsius scale associated with the metric system. In contrast to the English system, the metric system is neatly arranged according to the base-10 numerical framework: 10 millimeters to a centimeter, 100 centimeters to a meter, 1,000 meters to kilometer, and so on.

The difference between the pound and the kilogram aptly illustrates the reason scientists in general, and physicists in particular, prefer the metric system. A pound is a unit of weight, meaning that its value is entirely relative to the gravitational pull of the planet on which it is measured. A kilogram, on the other hand, is a unit of mass, and does not change throughout the universe. Though the basis for a kilogram may not ultimately be any more fundamental than that for a pound, it measures a quality that—unlike weight—does not vary according to frame of reference.

Frame of Reference in Classical Physics and Astronomy

Mass is a measure of inertia, the tendency of a body to maintain constant velocity. If an object is at rest, it tends to remain at rest, or if in motion, it tends to remain in motion unless acted upon by some outside force. This, as identified by the first law of motion, is inertia—and the greater the inertia, the greater the mass.

Physicists sometimes speak of an "inertial frame of reference," or one that has a constant velocity—that is, an unchanging speed and direction. Imagine if one were on a moving bus at constant velocity, regularly tossing a ball in the air and catching it. It would be no more difficult to catch the ball than if the bus were standing still, and indeed, there would be no way of determining, simply from the motion of the ball itself, that the bus was moving.

But what if the inertial frame of reference suddenly became a non-inertial frame of reference—in other words, what if the bus slammed on its brakes, thus changing its velocity? While the bus was moving forward, the ball was moving along with it, and hence, there was no relative motion between them. By stopping, the bus responded to an "outside" force—that is, its brakes. The ball, on the other hand, experienced that force indirectly. Hence, it would continue to move forward as before, in accordance with its own inertia—only now it would be in motion relative to the bus.

ASTRONOMY AND RELATIVE MOTION. The idea of relative motion plays a powerful role in astronomy. At every moment, Earth is turning on its axis at about 1,000 MPH (1,600 km/h) and hurtling along its orbital path around the Sun at the rate of 67,000 MPH (107,826 km/h.) The fastest any human being—that is, the astronauts taking part in the Apollo missions during the late 1960s—has traveled is about 30% of Earth's speed around the Sun.

Yet no one senses the speed of Earth's movement in the way that one senses the movement of a car—or indeed the way the astronauts perceived their speed, which was relative to the Moon and Earth. Of course, everyone experiences the results of Earth's movement—the change from night to day, the precession of the seasons—but no one experiences it directly. It is simply impossible, from the human frame of reference, to feel the movement of a body as large as Earth—not to mention larger progressions on the part of the Solar System and the universe.

FROM ASTRONOMY TO PHYSICS. The human body is in an inertial frame of reference with regard to Earth, and hence experiences no relative motion when Earth rotates or moves through space. In the same way, if one were traveling in a train alongside another train at constant velocity, it would be impossible to perceive that either train was actually moving—unless one referred to some fixed point, such as the trees or mountains in the background. Likewise, if two trains were sitting side by side, and one of them started to move, the relative motion might cause a person in the stationary train to believe that his or her train was the one moving.

For any measurement of velocity, and hence, of acceleration (a change in velocity), it is essential to establish a frame of reference. Velocity and acceleration, as well as inertia and mass, figured heavily in the work of Galileo Galilei (1564-1642) and Sir Isaac Newton (1642-1727), both of whom may be regarded as "founding fathers" of modern physics. Before Galileo, however, had come Nicholas Copernicus (1473-1543), the first modern astronomer to show that the Sun, and not Earth, is at the center of "the universe"—by which people of that time meant the Solar System.

In effect, Copernicus was saying that the frame of reference used by astronomers for millennia was incorrect: as long as they believed Earth to be the center, their calculations were bound to be wrong. Galileo and later Newton, through their studies in gravitation, were able to prove Copernicus's claim in terms of physics.

At the same time, without the understanding of a heliocentric (Sun-centered) universe that he inherited from Copernicus, it is doubtful that Newton could have developed his universal law of gravitation. If he had used Earth as the center-point for his calculations, the results would have been highly erratic, and no universal law would have emerged.

Relativity

For centuries, the model of the universe developed by Newton stood unchallenged, and even today it identifies the basic forces at work when speeds are well below that of the speed of light. However, with regard to the behavior of light itself—which travels at 186,000 mi (299,339 km) a second—Albert Einstein (1879-1955) began to observe phenomena that did not fit with Newtonian mechanics. The result of his studies was the Special Theory of Relativity, published in 1905, and the General Theory of Relativity, published a decade later. Together these altered humanity's view of the universe, and ultimately, of reality itself.

Einstein himself once offered this charming explanation of his epochal theory: "Put your hand on a hot stove for a minute, and it seems like an hour. Sit with a pretty girl for an hour, and it seems like a minute. That's relativity." Of course, relativity is not quite as simple as that—though the mathematics involved is no more challenging than that of a high-school algebra class. The difficulty lies in comprehending how things that seem impossible in the Newtonian universe become realities near the speed of light.

PLAYING TRICKS WITH TIME. An exhaustive explanation of relativity is far beyond the scope of the present discussion. What is important is the central precept: that no measurement of space or time is absolute, but depends on the relative motion of the observer (that is, the subject) and the observed (the object). Einstein further established that the movement of time itself is relative rather than absolute, a fact that would become apparent at speeds close to that of light. (His theory also showed that it is impossible to surpass that speed.)

Imagine traveling on a spaceship at nearly the speed of light while a friend remains stationary on Earth. Both on the spaceship and at the friend's house on Earth, there is a TV camera trained on a clock, and a signal relays the image from space to a TV monitor on Earth, and vice versa. What the TV monitor reveals is surprising: from your frame of reference on the spaceship, it seems that time is moving more slowly for your friend on Earth than for you. Your friend thinks exactly the same thing—only, from the friend's perspective, time on the spaceship is moving more slowly than time on Earth. How can this happen?

Again, a full explanation—requiring reference to formulae regarding time dilation, and so on—would be a rather involved undertaking. The short answer, however, is that which was stated above: no measurement of space or time is absolute, but each depends on the relative motion of the observer and the observed. Put another way, there is no such thing as absolute motion, either in the three dimensions of space, or in the fourth dimension identified by Einstein, time. All motion is relative to a frame of reference.

RELATIVITY AND ITS IMPLICATIONS. The ideas involved in relativity have been verified numerous times, and indeed the only reason why they seem so utterly foreign to most people is that humans are accustomed to living within the Newtonian framework. Einstein simply showed that there is no universal frame of reference, and like a true scientist, he drew his conclusions entirely from what the data suggested. He did not form an opinion, and only then seek the evidence to confirm it, nor did he seek to extend the laws of relativity into any realm beyond that which they described.

Yet British historian Paul Johnson, in his unorthodox history of the twentieth century, *Modern Times* (1983; revised 1992), maintained that a world disillusioned by World War I saw a moral dimension to relativity. Describing a set of tests regarding the behavior of the Sun's rays around the planet Mercury during an eclipse, the book begins with the sentence: "The modern world began on 29 May 1919, when photographs of a solar eclipse, taken on the Island of Principe off West Africa and at Sobral in Brazil, confirmed the truth of a new theory of the universe."

As Johnson went on to note, "...for most people, to whom Newtonian physics... were perfectly

KEY TERMS

ABSOLUTE: Fixed; not dependent on anything else. The value of 10 is absolute, relating to unchanging numerical principles; on the other hand, the value of 10 dollars is relative, reflecting the economy, inflation, buying power, exchange rates with other currencies, etc.

CALIBRATION: The process of checking and correcting the performance of a measuring instrument or device against a commonly accepted standard.

CARTESIAN COORDINATE SYSTEM: A method of specifying coordinates in relation to an x-axis, y-axis, and z-axis. The system is named after the French mathematician and philosopher René Descartes (1596-1650), who first described its principles, but it was developed greatly by French mathematician and philosopher Pierre de Fermat (1601-1665).

COEFFICIENT: A number that serves as a measure for some characteristic or property. A coefficient may also be a factor against which other values are multiplied to provide a desired result.

COORDINATE: A number or set of numbers used to specify the location of a point on a line, on a surface such as a plane, or in space.

FRAME OF REFERENCE: The perspective of a subject in observing an object.

OBJECT: Something that is perceived or observed by a subject.

RELATIVE: Dependent on something else for its value or for other identifying qualities. The fact that the United States has a constitution is an absolute, but the fact that it was ratified in 1787 is relative: that date has meaning only within the Western calendar.

SUBJECT: Something (usually a person) that perceives or observes an object and/or its behavior.

X-AXIS: The horizontal line of reference for points in the Cartesian coordinate system.

Y-AXIS: The vertical line of reference for points in the Cartesian coordinate system.

Z-AXIS: In a three-dimensional version of the Cartesian coordinate system, the z-axis is the line of reference for points in the third dimension. Typically the x-axis equates to "width," the y-axis to "height," and the z-axis to "depth"—though in fact length, width, and height are all relative to the observer's frame of reference.

comprehensible, relativity never became more than a vague source of unease. It was grasped that absolute time and absolute length had been dethroned.... All at once, nothing seemed certain in the spheres....At the beginning of the 1920s the belief began to circulate, for the first time at a popular level, that there were no longer any absolutes: of time and space, of good and evil, of knowledge, above all of value. Mistakenly but perhaps inevitably, relativity became confused with relativism."

Certainly many people agree that the twentieth century—an age that saw unprecedented mass murder under the dictatorships of Adolf Hitler and Josef Stalin, among others—was characterized by moral relativism, or the belief that there is no right or wrong. And just as Newton's discoveries helped usher in the Age of Reason, when thinkers believed it was possible to solve any problem through intellectual effort, it is quite plausible that Einstein's theory may have had this negative moral effect.

If so, this was certainly not Einstein's intention. Aside from the fact that, as stated, he did not set out to describe anything other than the physical behavior of objects, he continued to believe that there was no conflict between his ideas and a belief in an ordered universe: "Relativity," he once said, "teaches us the connection between the different descriptions of one and the same reality."

WHERE TO LEARN MORE

Beiser, Arthur. *Physics,* 5th ed. Reading, MA: Addison-Wesley, 1991.

Fleisher, Paul. *Relativity and Quantum Mechanics: Principles of Modern Physics.* Minneapolis, MN: Lerner Publications, 2002.

"*Frame of Reference*" (Web site). <http://www.physics.reading.ac.uk/units/flap/glossary/ff/frameref.html> (March 21, 2001).

"*Inertial Frame of Reference*" (Web site). <http://id.mind.net/~zona/mstm/physics/mechanics/framesOfReference /inertialFrame.html> (March 21, 2001).

Johnson, Paul. *Modern Times: The World from the Twenties to the Nineties.* Revised edition. New York: HarperPerennial, 1992.

King, Andrew. *Plotting Points and Position.* Illustrated by Tony Kenyon. Brookfield, CT: Copper Beech Books, 1998.

Parker, Steve. *Albert Einstein and Relativity.* New York: Chelsea House, 1995.

Robson, Pam. *Clocks, Scales, and Measurements.* New York: Gloucester Press, 1993.

Rutherford, F. James; Gerald Holton; and Fletcher G. Watson. *Project Physics.* New York: Holt, Rinehart, and Winston, 1981.

Swisher, Clarice. *Relativity: Opposing Viewpoints.* San Diego, CA: Greenhaven Press, 1990.

KINEMATICS AND DYNAMICS

CONCEPT

Webster's defines physics as "a science that deals with matter and energy and their interactions." Alternatively, physics can be described as the study of matter and motion, or of matter inmotion. Whatever the particulars of the definition, physics is among the most fundamental of disciplines, and hence, the rudiments of physics are among the most basic building blocks for thinking about the world. Foundational to an understanding of physics are kinematics, the explanation of how objects move, and dynamics, the study of why they move. Both are part of a larger branch of physics called mechanics, the study of bodies in motion. These are subjects that may sound abstract, but in fact, are limitless in their applications to real life.

HOW IT WORKS

The Place of Physics in the Sciences

Physics may be regarded as the queen of the sciences, not because it is "better" than chemistry or astronomy, but because it is the foundation on which all others are built. The internal and interpersonal behaviors that are the subject of the social sciences (psychology, anthropology, sociology, and so forth) could not exist without the biological framework that houses the human consciousness. Yet the human body and other elements studied by the biological and medical sciences exist within a larger environment, the framework for earth sciences, such as geology.

Earth sciences belong to a larger grouping of physical sciences, each more fundamental in concerns and broader in scope. Earth, after all, is but one corner of the realm studied by astronomy; and before a universe can even exist, there must be interactions of elements, the subject of chemistry. Yet even before chemicals can react, they have to do so within a physical framework—the realm of the most basic science—physics.

The Birth of Physics in Greece

THE FIRST HYPOTHESIS. Indeed, physics stands in relation to the sciences as philosophy does to thought itself: without philosophy to provide the concept of concepts, it would be impossible to develop a consistent worldview in which to test ideas. It is no accident, then, that the founder of the physical sciences was also the world's first philosopher, Thales (c. 625?-547? B.C.) of Miletus in Greek Asia Minor (now part of Turkey.) Prior to Thales's time, religious figures and mystics had made statements regarding ethics or the nature of deity, but none had attempted statements concerning the fundamental nature of reality.

For instance, the Bible offers a story of Earth's creation in the Book of Genesis which was well-suited to the understanding of people in the first millennium before Christ. But the writer of the biblical creation story made no attempt to explain how things came into being. He was concerned, rather, with showing that God had willed the existence of all physical reality by calling things into being—for example, by saying, "Let there be light."

Thales, on the other hand, made a genuine philosophical and scientific statement when he said that "Everything is water." This was the first hypothesis, a statement capable of being scientifically tested for accuracy. Thales's pronouncement did not mean he believed all things were necessarily made of water, literally. Rather, he appears to have been referring to a general tendency of movement: that the whole world is in a fluid state.

ATTEMPTING TO UNDERSTAND PHYSICAL REALITY. While we can respect Thales's statement for its truly earth-shattering implications, we may be tempted to read too much into it. Nonetheless, it is striking that he compared physical reality to water. On the one hand, there is the fact that water is essential to all life, and pervades Earth—but that is a subject more properly addressed by the realms of chemistry and the biological sciences. Perhaps of more interest to the physicist is the allusion to a fluid nature underlying all physical reality.

The physical realm is made of matter, which appears in four states: solid, liquid, gas, and plasma. The last of these is not the same as blood plasma: containing many ionized atoms or molecules which exhibit collective behavior, plasma is the substance from which stars, for instance, are composed. Though not plentiful on Earth, within the universe it may be the most common of all four states. Plasma is akin to gas, but different in molecular structure; the other three states differ at the molecular level as well.

Nonetheless, it is possible for a substance such as water—genuine H_2O, not the figurative water of Thales—to exist in liquid, gas, or solid form, and the dividing line between these is not always fixed. In fact, physicists have identified a phenomenon known as the triple point: at a certain temperature and pressure, a substance can be solid, liquid, and gas all at once!

The above statement shows just how challenging the study of physical reality can be, and indeed, these concepts would be far beyond the scope of Thales's imagination, had he been presented with them. Though he almost certainly deserves to be called a "genius," he lived in a world that viewed physical processes as a product of the gods' sometimes capricious will. The behavior of the tides, for instance, was attributed to Poseidon. Though Thales's statement began the process of digging humanity out from under the burden of superstition that had impeded scientific progress for centuries, the road forward would be a long one.

MATHEMATICS, MEASUREMENT, AND MATTER. In the two centuries after Thales's death, several other thinkers advanced understanding of physical reality in one way or another. Pythagoras (c. 580-c. 500 B.C.) taught that everything could be quantified, or related to numbers. Though he entangled this idea with mysticism and numerology, the concept itself influenced the idea that physical processes could be measured. Likewise, there were flaws at the heart of the paradoxes put forth by Zeno of Elea (c. 495-c. 430 B.C.), who set out to prove that motion was impossible—yet he was also the first thinker to analyze motion seriously.

In one of Zeno's paradoxes, he referred to an arrow being shot from a bow. At every moment of its flight, it could be said that the arrow was at rest within a space equal to its length. Though it would be some 2,500 years before slow-motion photography, in effect he was asking his listeners to imagine a snapshot of the arrow in flight. If it was at rest in that "snapshot," he asked, so to speak, and in every other possible "snapshot," when did the arrow actually move? These paradoxes were among the most perplexing questions of premodern times, and remain a subject of inquiry even today.

In fact, it seems that Zeno unwittingly (for there is no reason to believe that he deliberately deceived his listeners) inserted an error in his paradoxes by treating physical space as though it were composed of an infinite number of points. In the ideal world of geometric theory, a point takes up no space, and therefore it is correct to say that a line contains an infinite number of points; but this is not the case in the real world, where a "point" has some actual length. Hence, if the number of points on Earth were limitless, so too would be Earth itself.

Zeno's contemporary Leucippus (c. 480-c. 420 B.C.) and his student Democritus (c. 460-370 B.C.) proposed a new and highly advanced model for the tiniest point of physical space: the atom. It would be some 2,300 years, however, before physicists returned to the atomic model.

ARISTOTLE'S FLAWED PHYSICS

The study of matter and motion began to take shape with Aristotle (384-322 B.C.); yet, though his *Physics* helped establish a framework for the discipline, his errors are so profound that any praise must be qualified. Certainly, Aristotle was one of the world's greatest thinkers, who originated a set of formalized realms of study. However, in *Physics* he put forth an erroneous explanation of matter and motion that still prevailed in Europe twenty centuries later.

Actually, Aristotle's ideas disappeared in the late ancient period, as learning in general came to a virtual halt in Europe. That his writings—which on the whole did much more to advance the progress of science than to impede it—survived at all is a tribute to the brilliance of Arab, rather than European, civilization. Indeed, it was in the Arab world that the most important scientific work of the medieval period took place. Only after about 1200 did Aristotelian thinking once again enter Europe, where it replaced a crude jumble of superstitions that had been substituted for learning.

THE FOUR ELEMENTS. According to Aristotelian physics, all objects consisted, in varying degrees, of one or more elements: air, fire, water, and earth. In a tradition that went back to Thales, these elements were not necessarily pure: water in the everyday world was composed primarily of the element water, but also contained smaller amounts of the other elements. The planets beyond Earth were said to be made up of a "fifth element," or quintessence, of which little could be known.

The differing weights and behaviors of the elements governed the behavior of physical objects. Thus, water was lighter than earth, for instance, but heavier than air or fire. It was due to this difference in weight, Aristotle reasoned, that certain objects fall faster than others: a stone, for instance, because it is composed primarily of earth, will fall much faster than a leaf, which has much less earth in it.

Aristotle further defined "natural" motion as that which moved an object toward the center of the Earth, and "violent" motion as anything that propelled an object toward anything other than its "natural" destination. Hence, all horizontal or upward motion was "violent," and must be the direct result of a force. When the force was removed, the movement would end.

ARISTOTLE'S MODEL OF THE UNIVERSE. From the fact that Earth's center is the destination of all "natural" motion, it is easy to comprehend the Aristotelian cosmology, or model of the universe. Earth itself was in the center, with all other bodies (including the Sun)

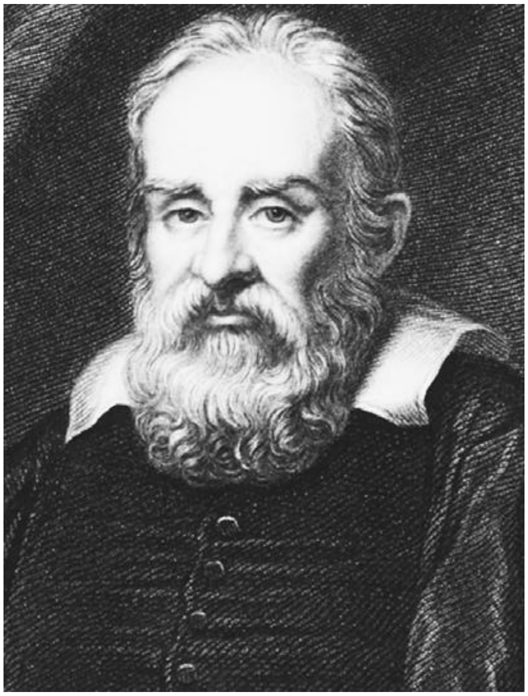
Aristotle. *(The Bettmann Archive. Reproduced by permission.)*

revolving around it. Though in constant movement, these heavenly bodies were always in their "natural" place, because they could only move on the firmly established—almost groove-like—paths of their orbits around Earth. This in turn meant that the physical properties of matter and motion on other planets were completely different from the laws that prevailed on Earth.

Of course, virtually every precept within the Aristotelian system is incorrect, and Aristotle compounded the influence of his errors by promoting a disdain for quantification. Specifically, he believed that mathematics had little value for describing physical processes in the real world, and relied instead on pure observation without attempts at measurement.

Moving Beyond Aristotle

Faulty as Aristotle's system was, however, it possessed great appeal because much of it seemed to fit with the evidence of the senses. It is not at all immediately apparent that Earth and the other planets revolve around the Sun, nor is it obvious that a stone and a leaf experience the same acceleration as they fall toward the ground. In fact, quite the opposite appears to be the case: as everyone knows, a stone falls faster than a leaf. Therefore, it would seem reasonable—on the



GALILEO. *(Archive Photos, Inc. Reproduced by permission.)*

surface of it, at least—to accept Aristotle's conclusion that this difference results purely from a difference in weight.

Today, of course, scientists—and indeed, even people without any specialized scientific knowledge—recognize the lack of merit in the Aristotelian system. The stone does fall faster than the leaf, but only because of air resistance, not weight. Hence, if they fell in a vacuum (a space otherwise entirely devoid of matter, including air), the two objects would fall at exactly the same rate.

As with a number of truths about matter and motion, this is not one that appears obvious, yet it has been demonstrated. To prove this highly nonintuitive hypothesis, however, required an approach quite different from Aristotle's—an approach that involved quantification and the separation of matter and motion into various components. This was the beginning of real progress in physics, and in a sense may be regarded as the true birth of the discipline. In the years that followed, understanding of physics would grow rapidly, thanks to advancements of many individuals; but their studies could not have been possible without the work of one extraordinary thinker who dared to question the Aristotelian model.

REAL-LIFE APPLICATIONS

KINEMATICS: HOW OBJECTS MOVE

By the sixteenth century, the Aristotelian worldview had become so deeply ingrained that few European thinkers would have considered the possibility that it could be challenged. Professors all over Europe taught Aristotle's precepts to their students, and in this regard the University of Pisa in Italy was no different. Yet from its classrooms would emerge a young man who not only questioned, but ultimately overturned the Aristotelian model: Galileo Galilei (1564-1642.)

Challenges to Aristotle had been slowly growing within the scientific communities of the Arab and later the European worlds during the preceding millennium. Yet the ideas that most influenced Galileo in his break with Aristotle came not from a physicist but from an astronomer, Nicolaus Copernicus (1473-1543.) It was Copernicus who made a case, based purely on astronomical observation, that the Sun and not Earth was at the center of the universe.

Galileo embraced this model of the cosmos, but was later forced to renounce it on orders from the pope in Rome. At that time, of course, the Catholic Church remained the single most powerful political entity in Europe, and its endorsement of Aristotelian views—which philosophers had long since reconciled with Christian ideas—is a measure of Aristotle's impact on thinking.

GALILEO'S REVOLUTION IN PHYSICS. After his censure by the Church, Galileo was placed under house arrest and was forbidden to study astronomy. Instead he turned to physics—where, ironically, he struck the blow that would destroy the bankrupt scientific system endorsed by Rome. In 1638, he published *Discourses and Mathematical Demonstrations Concerning Two New Sciences Pertaining to Mathematics and Local Motion,* a work usually referred to as *Two New Sciences.* In it, he laid the groundwork for physics by emphasizing a new method that included experimentation, demonstration, and quantification of results.

In this book—highly readable for a work of physics written in the seventeenth century—Galileo used a dialogue, an established format among philosophers and scientists of the past.

The character of Salviati argued for Galileo's ideas and Simplicio for those of Aristotle, while the genial Sagredo sat by and made occasional comments. Through Salviati, Galileo chose to challenge Aristotle on an issue that to most people at the time seemed relatively settled: the claim that objects fall at differing speeds according to their weight.

In order to proceed with his aim, Galileo had to introduce a number of innovations, and indeed, he established the subdiscipline of kinematics, or how objects move. Aristotle had indicated that when objects fall, they fall at the same rate from the moment they begin to fall until they reach their "natural" position. Galileo, on the other hand, suggested an aspect of motion, unknown at the time, that became an integral part of studies in physics: acceleration.

Scalars and Vectors

Even today, many people remain confused as to what acceleration is. Most assume that acceleration means only an increase in speed, but in fact this represents only one of several examples of acceleration. Acceleration is directly related to velocity, often mistakenly identified with speed.

In fact, speed is what scientists today would call a scalar quantity, or one that possesses magnitude but no specific direction. Speed is the rate at which the position of an object changes over a given period of time; thus people say "miles (or kilometers) per hour." A story problem concerning speed might state that "A train leaves New York City at a rate of 60 miles (96.6 km/h). How far will it have traveled in 73 minutes?"

Note that there is no reference to direction, whereas if the story problem concerned velocity—a vector, that is, a quantity involving both magnitude and direction—it would include some crucial qualifying phrase after "New York City": "for Boston," perhaps, or "northward." In practice, the difference between speed and velocity is nearly as large as that between a math problem and real life: few people think in terms of driving 60 miles, for instance, without also considering the direction they are traveling.

RESULTANTS. One can apply the same formula with velocity, though the process is more complicated. To obtain change in distance, one must add vectors, and this is best done by means of a diagram. You can draw each vector as an arrow on a graph, with the tail of each vector at the head of the previous one. Then it is possible to draw a vector from the tail of the first to the head of the last. This is the sum of the vectors, known as a resultant, which measures the net change.

Suppose, for instance, that a car travels east 4 mi (6.44 km), then due north 3 mi (4.83 km). This may be drawn on a graph with four units along the x axis, then 3 units along the y axis, making two sides of a triangle. The number of sides to the resulting shape is always one more than the number of vectors being added; the final side is the resultant. From the tail of the first segment, a diagonal line drawn to the head of the last will yield a measurement of 5 units—the resultant, which in this case would be equal to 5 mi (8 km) in a northeasterly direction.

VELOCITY AND ACCELERATION. The directional component of velocity makes it possible to consider forms of motion other than linear, or straight-line, movement. Principal among these is circular, or rotational motion, in which an object continually changes direction and thus, velocity. Also significant is projectile motion, in which an object is thrown, shot, or hurled, describing a path that is a combination of horizontal and vertical components.

Furthermore, velocity is a key component in acceleration, which is defined as a change in velocity. Hence, acceleration can mean one of five things: an increase in speed with no change in direction (the popular, but incorrect, definition of the overall concept); a decrease in speed with no change in direction; a decrease or increase of speed with a change in direction; or a change in direction with no change in speed. If a car speeds up or slows down while traveling in a straight line, it experiences acceleration. So too does an object moving in rotational motion, even if its speed does not change, because its direction will change continuously.

Dynamics: Why Objects Move

GALILEO'S TEST. To return to Galileo, he was concerned primarily with a specific form of acceleration, that which occurs due to the force of gravity. Aristotle had provided an explanation of gravity—if a highly flawed one—with his claim that objects fall to their "natural" position; Galileo set out to develop the first truly scientific explanation concerning how objects fall to the ground.

According to Galileo's predictions, two metal balls of differing sizes would fall with the same rate of acceleration. To test his hypotheses, however, he could not simply drop two balls from a rooftop—or have someone else do so while he stood on the ground—and measure their rate of fall. Objects fall too fast, and lacking sophisticated equipment available to scientists today, he had to find another means of showing the rate at which they fell.

This he did by resorting to a method Aristotle had shunned: the use of mathematics as a means of modeling the behavior of objects. This is such a deeply ingrained aspect of science today that it is hard to imagine a time when anyone would have questioned it, and that very fact is a tribute to Galileo's achievement. Since he could not measure speed, he set out to find an equation relating total distance to total time. Through a detailed series of steps, Galileo discovered that in uniform or constant acceleration from rest—that is, the acceleration he believed an object experiences due to gravity—there is a proportional relationship between distance and time.

With this mathematical model, Galileo could demonstrate uniform acceleration. He did this by using an experimental model for which observation was easier than in the case of two falling bodies: an inclined plane, down which he rolled a perfectly round ball. This allowed him to extrapolate that in free fall, though velocity was greater, the same proportions still applied and therefore, acceleration was constant.

POINTING THE WAY TOWARD NEWTON. The effects of Galileo's system were enormous: he demonstrated mathematically that acceleration is constant, and established a method of hypothesis and experiment that became the basis of subsequent scientific investigation. He did not, however, attempt to calculate a figure for the acceleration of bodies in free fall; nor did he attempt to explain the overall principle of gravity, or indeed why objects move as they do—the focus of a subdiscipline known as dynamics.

At the end of *Two New Sciences,* Sagredo offered a hopeful prediction: "I really believe that... the principles which are set forth in this little treatise will, when taken up by speculative minds, lead to another more remarkable result...." This prediction would come true with the work of a man who, because he lived in a somewhat more enlightened time—and because he lived in England, where the pope had no power—was free to explore the implications of his physical studies without fear of Rome's intervention. Born in the very year Galileo died, his name was Sir Isaac Newton (1642-1727.)

NEWTON'S THREE LAWS OF MOTION. In discussing the movement of the planets, Galileo had coined the term inertia to describe the tendency of an object in motion to remain in motion, and an object at rest to remain at rest. This idea would be the starting point of Newton's three laws of motion, and Newton would greatly expand on the concept of inertia.

The three laws themselves are so significant to the understanding of physics that they are treated separately elsewhere in this volume; here they are considered primarily in terms of their implications regarding the larger topic of matter and motion.

Introduced by Newton in his *Principia* (1687), the three laws are:

- First law of motion: An object at rest will remain at rest, and an object in motion will remain in motion, at a constant velocity unless or until outside forces act upon it.
- Second law of motion: The net force acting upon an object is a product of its mass multiplied by its acceleration.
- Third law of motion: When one object exerts a force on another, the second object exerts on the first a force equal in magnitude but opposite in direction.

These laws made final the break with Aristotle's system. In place of "natural" motion, Newton presented the concept of motion at a uniform velocity—whether that velocity be a state of rest or of uniform motion. Indeed, the closest thing to "natural" motion (that is, true "natural" motion) is the behavior of objects in outer space. There, free from friction and away from the gravitational pull of Earth or other bodies, an object set in motion will remain in motion forever due to its own inertia. It follows from this observation, incidentally, that Newton's laws were and are universal, thus debunking the old myth that the physical properties of realms outside Earth are fundamentally different from those of Earth itself.

MASS AND GRAVITATIONAL ACCELERATION. The first law establishes the principle of inertia, and the second law makes reference to the means by which inertia is measured: mass, or the resistance of an object to a

KEY TERMS

ACCELERATION: A change in velocity.

DYNAMICS: The study of why objects move as they do; compare with kinematics.

FORCE: The product of mass multiplied by acceleration.

HYPOTHESIS: A statement capable of being scientifically tested for accuracy.

INERTIA: The tendency of an object in motion to remain in motion, and of an object at rest to remain at rest.

KINEMATICS: The study of how objects move; compare with dynamics.

MASS: A measure of inertia, indicating the resistance of an object to a change in its motion—including a change in velocity.

MATTER: The material of physical reality. There are four basic states of matter: solid, liquid, gas, and plasma.

MECHANICS: The study of bodies in motion.

RESULTANT: The sum of two or more vectors, which measures the net change in distance and direction.

SCALAR: A quantity that possesses only magnitude, with no specific direction. Mass, time, and speed are all scalars. The opposite of a scalar is a vector.

SPEED: The rate at which the position of an object changes over a given period of time.

VACUUM: Space entirely devoid of matter, including air.

VECTOR: A quantity that possesses both magnitude and direction. Velocity, acceleration, and weight (which involves the downward acceleration due to gravity) are examples of vectors. Its opposite is a scalar.

VELOCITY: The speed of an object in a particular direction.

WEIGHT: A measure of the gravitational force on an object; the product of mass multiplied by the acceleration due to gravity. (The latter is equal to 32 ft or 9.8 m per second per second, or 32 ft/9.8 m per second squared.)

change in its motion—including a change in velocity. Mass is one of the most fundamental notions in the world of physics, and it too is the subject of a popular misconception—one which confuses it with weight. In fact, weight is a force, equal to mass multiplied by the acceleration due to gravity.

It was Newton, through a complicated series of steps he explained in his *Principia,* who made possible the calculation of that acceleration—an act of quantification that had eluded Galileo. The figure most often used for gravitational acceleration at sea level is 32 ft (9.8 m) per second squared. This means that in the first second, an object falls at a velocity of 32 ft per second, but its velocity is also increasing at a rate of 32 ft per second per second. Hence, after 2 seconds, its velocity will be 64 ft (per second; after 3 seconds 96 ft per second, and so on.

Mass does not vary anywhere in the universe, whereas weight changes with any change in the gravitational field. When United States astronaut Neil Armstrong planted the American flag on the Moon in 1969, the flagpole (and indeed Armstrong himself) weighed much less than on Earth. Yet it would have required exactly the same amount of force to move the pole (or, again, Armstrong) from side to side as it would have on Earth, because their mass and therefore their inertia had not changed.

Beyond Mechanics

The implications of Newton's three laws go far beyond what has been described here; but again, these laws, as well as gravity itself, receive a much more thorough treatment elsewhere in this volume. What is important in this context is the gradually unfolding understanding of matter and motion that formed the basis for the study of physics today.

After Newton came the Swiss mathematician and physicist Daniel Bernoulli (1700-1782), who pioneered another subdiscipline, fluid dynamics, which encompasses the behavior of liquids and gases in contact with solid objects. Air itself is an example of a fluid, in the scientific sense of the term. Through studies in fluid dynamics, it became possible to explain the principles of air resistance that cause a leaf to fall more slowly than a stone—even though the two are subject to exactly the same gravitational acceleration, and would fall at the same speed in a vacuum.

EXTENDING THE REALM OF PHYSICAL STUDY. The work of Galileo, Newton, and Bernoulli fit within one of five major divisions of classical physics: mechanics, or the study of matter, motion, and forces. The other principal divisions are acoustics, or studies in sound; optics, the study of light; thermodynamics, or investigations regarding the relationships between heat and other varieties of energy; and electricity and magnetism. (These subjects, and subdivisions within them, also receive extensive treatment elsewhere in this book.)

Newton identified one type of force, gravitation, but in the period leading up to the time of Scottish physicist James Clerk Maxwell (1831-1879), scientists gradually became aware of a new fundamental interaction in the universe. Building on studies of numerous scientists, Maxwell hypothesized that electricity and magnetism are in fact differing manifestations of a second variety of force, electromagnetism.

MODERN PHYSICS. The term classical physics, used above, refers to the subjects of study from Galileo's time through the end of the nineteenth century. Classical physics deals primarily with subjects that can be discerned by the senses, and addressed processes that could be observed on a large scale. By contrast, modern physics, which had its beginnings with the work of Max Planck (1858-1947), Albert Einstein (1879-1955), Niels Bohr (1885-1962), and others at the beginning of the twentieth century, addresses quite a different set of topics.

Modern physics is concerned primarily with the behavior of matter at the molecular, atomic, or subatomic level, and thus its truths cannot be grasped with the aid of the senses. Nor is classical physics much help in understanding modern physics. The latter, in fact, recognizes two forces unknown to classical physicists: weak nuclear force, which causes the decay of some subatomic particles, and strong nuclear force, which binds the nuclei of atoms with a force 1 trillion (10^{12}) times as great as that of the weak nuclear force.

Things happen in the realm of modern physics that would have been inconceivable to classical physicists. For instance, according to quantum mechanics—first developed by Planck—it is not possible to make a measurement without affecting the object (e.g., an electron) being measured. Yet even atomic, nuclear, and particle physics can be understood in terms of their effects on the world of experience: challenging as these subjects are, they still concern—though within a much more complex framework—the physical fundamentals of matter and motion.

WHERE TO LEARN MORE

Ballard, Carol. *How Do We Move?* Austin, TX: Raintree Steck-Vaughn, 1998.

Beiser, Arthur. *Physics,* 5th ed. Reading, MA: Addison-Wesley, 1991.

Fleisher, Paul. *Objects in Motion: Principles of Classical Mechanics.* Minneapolis, MN: Lerner Publications, 2002.

Hewitt, Sally. *Forces Around Us.* New York: Children's Press, 1998.

Measure for Measure: Sites That Do the Work for You (Web site). <http://www.wolinskyweb.com/measure.html> (March 7, 2001).

Motion, Energy, and Simple Machines (Web site). <http://www.necc.mass.edu/MRVIS/MR3_13/start.html> (March 7, 2001).

Physlink.com (Web site). <http://www.physlink.com> (March 7, 2001).

Rutherford, F. James; Gerald Holton; and Fletcher G. Watson. *Project Physics.* New York: Holt, Rinehart, and Winston, 1981.

Wilson, Jerry D. *Physics: Concepts and Applications,* second edition. Lexington, MA: D. C. Heath, 1981.

DENSITY AND VOLUME

CONCEPT

Density and volume are simple topics, yet in order to work within any of the hard sciences, it is essential to understand these two types of measurement, as well as the fundamental quantity involved in conversions between them—mass. Measuring density makes it possible to distinguish between real gold and fake gold, and may also give an astronomer an important clue regarding the internal composition of a planet.

HOW IT WORKS

There are four fundamental standards by which most qualities in the physical world can be measured: length, mass, time, and electric current. The volume of a cube, for instance, is a unit of length cubed: the length is multiplied by the width and multiplied by the height. Width and height, however, are not distinct standards of measurement: they are simply versions of length, distinguished by their orientation. Whereas length is typically understood as a distance along an x-axis in one-dimensional space, width adds a second dimension, and height a third.

Of particular concern within this essay are length and mass, since volume is measured in terms of length, and density in terms of the ratio between mass and volume. Elsewhere in this book, the distinction between mass and weight has been presented numerous times from the standpoint of a person whose mass and weight are measured on Earth, and again on the Moon. Mass, of course, does not change, whereas weight does, due to the difference in gravitational force exerted by Earth as compared with that of its satellite, the Moon. But consider instead the role of the fundamental quality, mass, in determining this significantly less fundamental property of weight.

According to the second law of motion, weight is a force equal to mass multiplied by acceleration. Acceleration, in turn, is equal to change in velocity divided by change in time. Velocity, in turn, is equal to distance (a form of length) divided by time. If one were to express weight in terms of *l, t,* and *m,* with these representing, respectively, the fundamental properties of length, time, and mass, it would be expressed as

$$\frac{M \cdot D}{t^2}$$

—clearly, a much more complicated formula than that of mass!

MASS

So what is mass? Again, the second law of motion, derived by Sir Isaac Newton (1642-1727), is the key: mass is the ratio of force to acceleration. This topic, too, is discussed in numerous places throughout this book; what is actually of interest here is a less precise identification of mass, also made by Newton.

Before formulating his laws of motion, Newton had used a working definition of mass as the quantity of matter an object possesses. This is not of much value for making calculations or measurements, unlike the definition in the second law. Nonetheless, it serves as a useful reminder of matter's role in the formula for density.

Matter can be defined as a physical substance not only having mass, but occupying space. It is composed of atoms (or in the case of subatomic particles, it is part of an atom), and is

HOW DOES A GIGANTIC STEEL SHIP, SUCH AS THE SUPERTANKER PICTURED HERE, STAY AFLOAT, EVEN THOUGH IT HAS A WEIGHT DENSITY FAR GREATER THAN THE WATER BELOW IT? THE ANSWER LIES IN ITS CURVED HULL, WHICH CONTAINS A LARGE AMOUNT OF OPEN SPACE AND ALLOWS THE SHIP TO SPREAD ITS AVERAGE DENSITY TO A LOWER LEVEL THAN THE WATER. (*Photograph by Vince Streano/Corbis. Reproduced by permission.*)

convertible with energy. The form or state of matter itself is not important: on Earth it is primarily observed as a solid, liquid, or gas, but it can also be found (particularly in other parts of the universe) in a fourth state, plasma.

Yet there are considerable differences among types of matter—among various elements and states of matter. This is apparent if one imagines three gallon jugs, one containing water, the second containing helium, and the third containing iron filings. The volume of each is the same, but obviously, the mass is quite different.

The reason, of course, is that at a molecular level, there is a difference in mass between the compound H_2O and the elements helium and iron. In the case of helium, the second-lightest of all elements after hydrogen, it would take a great deal of helium for its mass to equal that of iron. In fact, it would take more than 43,000 gallons of helium to equal the mass of the iron in one gallon jug!

DENSITY

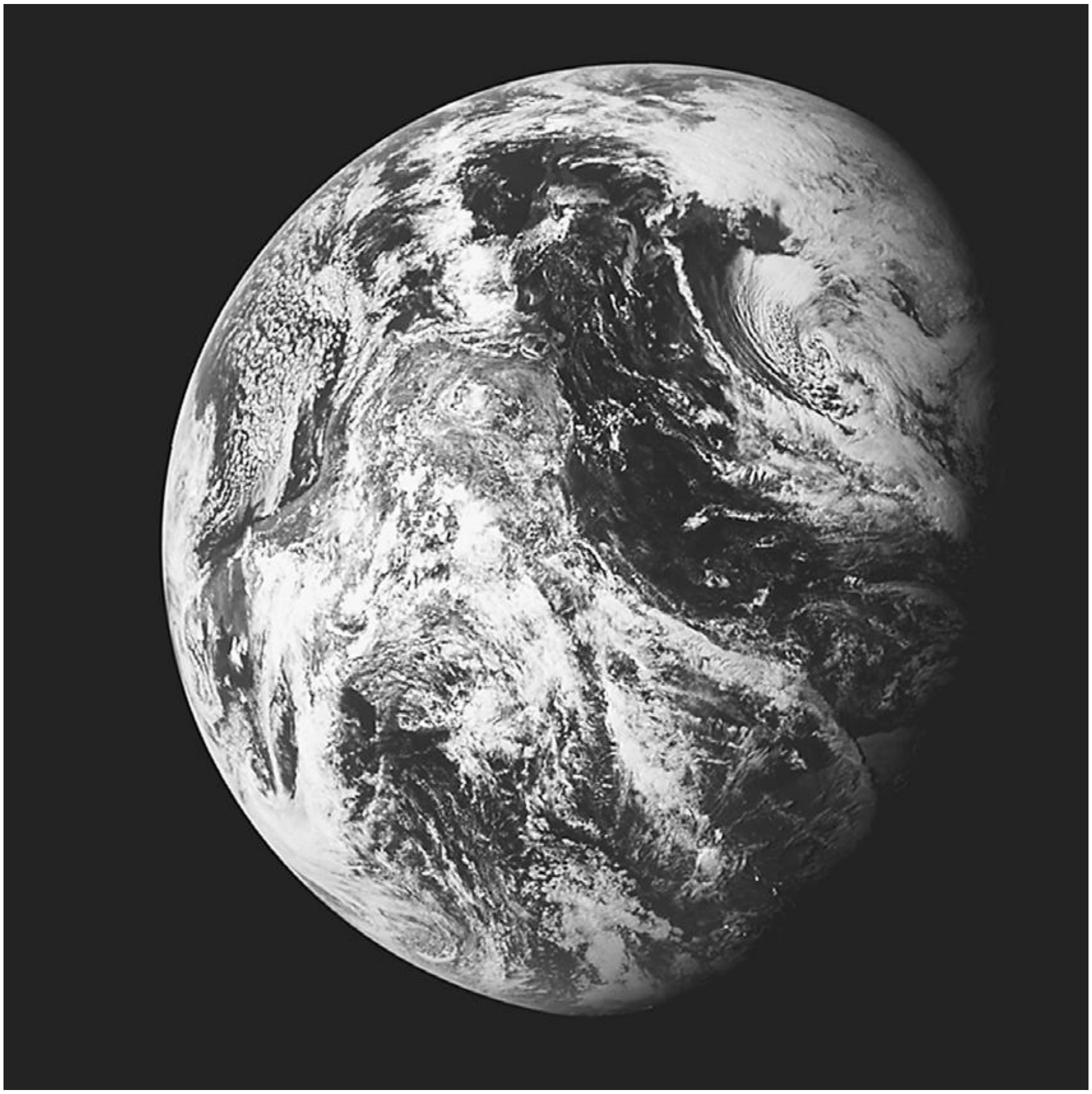
Rather than comparing differences in molecular mass among the three substances, it is easier to analyze them in terms of density, or mass divided by volume. It so happens that the three items represent the three states of matter on Earth: liquid (water), solid (iron), and gas (helium). For the most part, solids tend to be denser than liquids, and liquids denser than gasses.

One of the interesting things about density, as distinguished from mass and volume, is that it has nothing to do with the amount of material. A kilogram of iron differs from 10 kilograms of iron both in mass and volume, but the density of both samples is the same. Indeed, as discussed below, the known densities of various materials make it possible to determine whether a sample of that material is genuine.

VOLUME

Mass, because of its fundamental nature, is sometimes hard to comprehend, and density requires an explanation in terms of mass and volume. Volume, on the other hand, appears to be quite straightforward—and it is, when one is describing a solid of regular shape. In other situations, however, volume is more complicated.

As noted earlier, the volume of a cube can be obtained simply by multiplying length by width by height. There are other means for measuring



SINCE SCIENTISTS KNOW EARTH'S MASS AS WELL AS ITS VOLUME, THEY ARE EASILY ABLE TO COMPUTE ITS DENSITY—APPROXIMATELY 5 G/CM^3. *(Corbis. Reproduced by permission.)*

the volume of other straight-sided objects, such as a pyramid. That formula applies, indeed, for any polyhedron (a three-dimensional closed solid bounded by a set number of plane figures) that constitutes a modified cube in which the lengths of the three dimensions are unequal—that is, an oblong shape.

For a cylinder or sphere, volume measurements can be obtained by applying formulae involving radius (r) and the constant π, roughly equal to 3.14. The formula for volume of a cylinder is $V = \pi r^2 h$, where h is the height. A sphere's volume can be obtained by the formula $(4/3)\pi r^3$. Even the volume of a cone can be easily calculated: it is one-third that of a cylinder of equal base and height.

REAL-LIFE APPLICATIONS

Measuring Volume

What about the volume of a solid that is irregular in shape? Some irregularly shaped objects, such as a scooter, which consists primarily of one round wheel and a number of oblong shapes, can be measured by separating them into regular shapes. Calculus may be employed with more complex problems to obtain the volume of an irregular shape—but the most basic method is simply to immerse the object in water. This procedure involves measuring the volume of the water before and after immersion, and calculating the difference. Of course, the object being measured cannot be water-soluble; if it is, its volume must be measured in a non-water-based liquid such as alcohol.

Measuring liquid volumes is easy, given the fact that liquids have no definite shape, and will simply take the shape of the container in which they are placed. Gases are similar to liquids in the sense that they expand to fit their container; however, measurement of gas volume is a more involved process than that used to measure either liquids or solids, because gases are highly responsive to changes in temperature and pressure.

If the temperature of water is raised from its freezing point to its boiling point (32° to 212°F or 0 to 100°C), its volume will increase by only 2%. If its pressure is doubled from 1 atm (defined as normal air pressure at sea level—14.7 pounds-per-square-inch or 1.013×10^5 Pa) to 2 atm, volume will decrease by only 0.01%.

Yet, if air were heated from 32° to 212°F, its volume would increase by 37%; and if its pressure were doubled from 1 atm to 2, its volume would decrease by 50%. Not only do gases respond dramatically to changes in temperature and pressure, but also, gas molecules tend to be non-attractive toward one another—that is, they do not tend to stick together. Hence, the concept of "volume" involving gas is essentially meaningless, unless its temperature and pressure are known.

Buoyancy: Volume and Density

Consider again the description above, of an object with irregular shape whose volume is measured by immersion in water. This is not the only interesting use of water and solids when dealing with volume and density. Particularly intriguing is the concept of buoyancy expressed in Archimedes's principle.

More than twenty-two centuries ago, the Greek mathematician, physicist, and inventor Archimedes (c. 287-212 B.C.) received orders from the king of his hometown—Syracuse, a Greek colony in Sicily—to weigh the gold in the royal crown. According to legend, it was while bathing that Archimedes discovered the principle that is today named after him. He was so excited, legend maintains, that he jumped out of his bath and ran naked through the streets of Syracuse shouting "Eureka!" (I have found it).

What Archimedes had discovered was, in short, the reason why ships float: because the buoyant, or lifting, force of an object immersed in fluid is equal to the weight of the fluid displaced by the object.

HOW A STEEL SHIP FLOATS ON WATER. Today most ships are made of steel, and therefore, it is even harder to understand why an aircraft carrier weighing many thousands of tons can float. After all, steel has a weight density (the preferred method for measuring density according to the British system of measures) of 480 pounds per cubic foot, and a density of 7,800 kilograms-per-cubic-meter. By contrast, sea water has a weight density of 64 pounds per cubic foot, and a density of 1,030 kilograms-per-cubic-meter.

This difference in density should mean that the carrier would sink like a stone—and indeed it would, if all the steel in it were hammered flat. As it is, the hull of the carrier (or indeed of any seaworthy ship) is designed to displace or move a quantity of water whose weight is greater than that of the vessel itself. The weight of the displaced water—that is, its mass multiplied by the downward acceleration due to gravity—is equal to the buoyant force that the ocean exerts on the ship. If the ship weighs less than the water it displaces, it will float; but if it weighs more, it will sink.

Put another way, when the ship is placed in the water, it displaces a certain quantity of water whose weight can be expressed in terms of Vdg—volume multiplied by density multiplied by the downward acceleration due to gravity. The density of sea water is a known figure, as is g (32 ft or 9.8 m/sec^2); thus the only variable for the water displaced is its volume.

For the buoyant force on the ship, g will of course be the same, and the value of V will be the same as for the water. In order for the ship to float, then, its density must be much less than that of the water it has displaced. This can be achieved by designing the ship in order to maximize displacement. The steel is spread over as large an area as possible, and the curved hull, when seen in cross section, contains a relatively large area of open space. Obviously, the density of this space is much less than that of water; thus, the average density of the ship is greatly reduced, which enables it to float.

KEY TERMS

ARCHIMEDES'S PRINCIPLE: A rule of physics which holds that the buoyant force of an object immersed in fluid is equal to the weight of the fluid displaced by the object. It is named after the Greek mathematician, physicist, and inventor Archimedes (c. 287-212 B.C.), who first identified it.

BUOYANCY: The tendency of an object immersed in a fluid to float. This can be explained by Archimedes's principle.

DENSITY: The ratio of mass to volume—in other words, the amount of matter within a given area.

MASS: According to the second law of motion, mass is the ratio of force to acceleration. Mass may likewise be defined, though much less precisely, as the amount of matter an object contains. Mass is also the product of volume multiplied by density.

MATTER: Physical substance that occupies space, has mass, is composed of atoms (or in the case of subatomic particles, is part of an atom), and is convertible into energy.

SPECIFIC GRAVITY: The density of an object or substance relative to the density of water; or more generally, the ratio between the densities of two objects or substances.

VOLUME: The amount of three-dimensional space an object occupies. Volume is usually expressed in cubic units of length.

WEIGHT DENSITY: The proper term for density within the British system of weights and measures. The pound is a unit of weight rather than of mass, and thus British units of density are usually rendered in terms of weight density—that is, pounds-per-cubic-foot. By contrast, the metric or international units measure mass density (referred to simply as "density"), which is rendered in terms of kilograms-per-cubic-meter, or grams-per-cubic-centimeter.

Comparing Densities

As noted several times, the densities of numerous materials are known quantities, and can be easily compared. Some examples of density, all expressed in terms of kilograms per cubic meter, are:

- Hydrogen: 0.09 kg/m^3
- Air: 1.3 kg/m^3
- Oak: 720 kg/m^3
- Ethyl alcohol: 790 kg/m^3
- Ice: 920 kg/m^3
- Pure water: 1,000 kg/m^3
- Concrete: 2,300 kg/m^3
- Iron and steel: 7,800 kg/m^3
- Lead: 11,000 kg/m^3
- Gold: 19,000 kg/m^3

Note that pure water (as opposed to sea water, which is 3% denser) has a density of 1,000 kilograms per cubic meter, or 1 gram per cubic centimeter. This value is approximate; however, at a temperature of 39.2°F (4°C) and under normal atmospheric pressure, it is exact, and so, water is a useful standard for measuring the specific gravity of other substances.

SPECIFIC GRAVITY AND THE DENSITIES OF PLANETS. Specific gravity is the ratio between the densities of two objects or substances, and it is expressed as a number without units of measure. Due to the value of 1 g/cm^3 for water, it is easy to determine the specific gravity of a given substance, which will have the same number value as its density. For example, the specific gravity of concrete, which has a density of 2.3 g/cm^3, is 2.3. The specific gravities of gases are usually determined in comparison to the specific gravity of dry air.

Most rocks near the surface of Earth have a specific gravity of somewhere between 2 and 3, while the specific gravity of the planet itself is about 5. How do scientists know that the density of Earth is around 5 g/cm^3? The computation is fairly simple, given the fact that the mass and volume of the planet are known. And given the fact that most of what lies close to Earth's surface—sea water, soil, rocks—has a specific gravity well below 5, it is clear that Earth's interior must contain high-density materials, such as nickel or iron. In the same way, calculations regarding the density of other objects in the Solar System provide a clue as to their interior composition.

ALL THAT GLITTERS. Closer to home, a comparison of density makes it possible to determine whether a piece of jewelry alleged to be solid gold is really genuine. To determine the answer, one must drop it in a beaker of water with graduated units of measure clearly marked. (Here, figures are given in cubic centimeters, since these are easiest to use in this context.)

Suppose the item has a mass of 10 grams. The density of gold is 19 g/cm^3, and since $V = m/d$ = 10/19, the volume of water displaced by the gold should be 0.53 cm^3. Suppose that instead, the item displaced 0.91 cm^3 of water. Clearly, it is not gold, but what is it?

Given the figures for mass and volume, its density would be equal to m/V = 10/0.91 = 11 g/cm^3—which happens to be the density of lead. If on the other hand the amount of water displaced were somewhere between the values for pure gold and pure lead, one could calculate what portion of the item was gold and which lead. It is possible, of course, that it could contain some other metal, but given the high specific gravity of lead, and the fact that its density is relatively close to that of gold, lead is a favorite gold substitute among jewelry counterfeiters.

WHERE TO LEARN MORE

Beiser, Arthur. *Physics,* 5th ed. Reading, MA: Addison-Wesley, 1991.

Chahrour, Janet. *Flash! Bang! Pop! Fizz!: Exciting Science for Curious Minds.* Illustrated by Ann Humphrey Williams. Hauppauge, N.Y.: Barron's, 2000.

"*Density and Specific Gravity*" (Web site). <http://www.tpub.com/fluid/ch1e.htm> (March 27, 2001).

"*Density, Volume, and Cola*" (Web site). <http://student.biology.arizona.edu/sciconn/density/density_coke.html> (March 27, 2001).

"*The Mass Volume Density Challenge*" (Web site). <http://science-math-technology.com/mass_volume_density.html> (March 27, 2001).

"*Metric Density and Specific Gravity*" (Web site). <http://www.essex1.com/people/speer/density.html> (March 27, 2001).

"*Mineral Properties: Specific Gravity*" *The Mineral and Gemstone Kingdom* (Web site). <http://www.minerals.net/resource/property/sg.htm> (March 27, 2001).

Robson, Pam. *Clocks, Scales and Measurements.* New York: Gloucester Press, 1993.

"*Volume, Mass, and Density*" (Web site). <http://www.nyu.edu/pages/mathmol/modules/water/density_intro.html> (March 27, 2001).

Willis, Shirley. *Tell Me How Ships Float.* Illustrated by the author. New York: Franklin Watts, 1999.

CONSERVATION LAWS

CONCEPT

The term "conservation laws" might sound at first like a body of legal statutes geared toward protecting the environment. In physics, however, the term refers to a set of principles describing certain aspects of the physical universe that are preserved throughout any number of reactions and interactions. Among the properties conserved are energy, linear momentum, angular momentum, and electrical charge. (Mass, too, is conserved, though only in situations well below the speed of light.) The conservation of these properties can be illustrated by examples as diverse as dropping a ball (energy); the motion of a skater spinning on ice (angular momentum); and the recoil of a rifle (linear momentum).

HOW IT WORKS

The conservation laws describe physical properties that remain constant throughout the various processes that occur in the physical world. In physics, "to conserve" something means "to result in no net loss of" that particular component. For each such component, the input is the same as the output: if one puts a certain amount of energy into a physical system, the energy that results from that system will be the same as the energy put into it.

The energy may, however, change forms. In addition, the operations of the conservation laws are—on Earth, at least—usually affected by a number of other forces, such as gravity, friction, and air resistance. The effects of these forces, combined with the changes in form that take place within a given conserved property, sometimes make it difficult to perceive the working of the conservation laws. It was stated above that the resulting energy of a physical system will be the same as the energy that was introduced to it. Note, however, that the usable energy output of a system will not be equal to the energy input. This is simply impossible, due to the factors mentioned above—particularly friction.

When one puts gasoline into a motor, for instance, the energy that the motor puts out will never be as great as the energy contained in the gasoline, because part of the input energy is expended in the operation of the motor itself. Similarly, the angular momentum of a skater on ice will ultimately be dissipated by the resistant force of friction, just as that of a Frisbee thrown through the air is opposed both by gravity and air resistance—itself a specific form of friction.

In each of these cases, however, the property is still conserved, even if it does not seem so to the unaided senses of the observer. Because the motor has a usable energy output less than the input, it seems as though energy has been lost. In fact, however, the energy has only changed forms, and some of it has been diverted to areas other than the desired output. (Both the noise and the heat of the motor, for instance, represent uses of energy that are typically considered undesirable.) Thus, upon closer study of the motor—itself an example of a system—it becomes clear that the resulting energy, if not the desired usable output, is the same as the energy input.

As for the angular momentum examples in which friction, or air resistance, plays a part, here too (despite all apparent evidence to the contrary) the property is conserved. This is easier to understand if one imagines an object spinning in outer space, free from the opposing force of friction. Thanks to the conservation of angular

AS THIS HUNTER FIRES HIS RIFLE, THE RIFLE PRODUCES A BACKWARD "KICK" AGAINST HIS SHOULDER. THIS KICK, WITH A VELOCITY IN THE OPPOSITE DIRECTION OF THE BULLET'S TRAJECTORY, HAS A MOMENTUM EXACTLY THE SAME AS THAT OF THE BULLET ITSELF: HENCE MOMENTUM IS CONSERVED. *(Photograph by Tony Arruza/Corbis. Reproduced by permission.)*

momentum, an object set into rotation in space will continue to spin indefinitely. Thus, if an astronaut in the 1960s, on a spacewalk from his capsule, had set a screwdriver spinning in the emptiness of the exosphere, the screwdriver would still be spinning today!

ENERGY AND MASS

Among the most fundamental statements in all of science is the conservation of energy: a system isolated from all outside factors will maintain the same total amount of energy, even though energy transformations from one form or another take place.

Energy is manifested in many varieties, including thermal, electromagnetic, sound, chemical, and nuclear energy, but all these are merely reflections of three basic types of energy. There is potential energy, which an object possesses by virtue of its position; kinetic energy, which it possesses by virtue of its motion; and rest energy, which it possesses by virtue of its mass.

The last of these three will be discussed in the context of the relationship between energy and mass; at present the concern is with potential and kinetic energy. Every system possesses a certain quantity of both, and the sum of its potential and kinetic energy is known as mechanical energy. The mechanical energy within a system does not change, but the relative values of potential and kinetic energy may be altered.

A SIMPLE EXAMPLE OF MECHANICAL ENERGY. If one held a baseball at the top of a tall building, it would have a certain amount of potential energy. Once it was dropped, it would immediately begin losing potential energy and gaining kinetic energy proportional to the potential energy it lost. The relationship between the two forms, in fact, is inverse: as the value of one variable decreases, that of the other increases in exact proportion.

The ball cannot keep falling forever, losing potential energy and gaining kinetic energy. In fact, it can never gain an amount of kinetic energy greater than the potential energy it possessed in the first place. At the moment before it hits the ground, the ball's kinetic energy is equal to the potential energy it possessed at the top of the building. Correspondingly, its potential energy is zero—the same amount of kinetic energy it possessed before it was dropped.

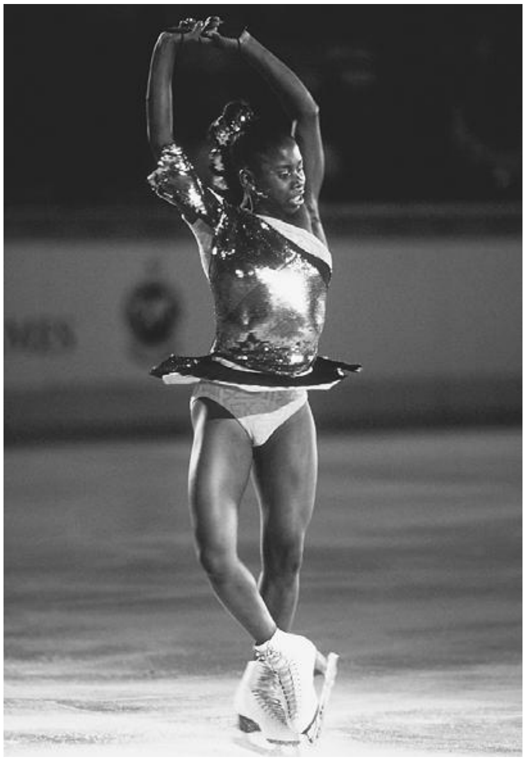
Then, as the ball hits the ground, the energy is dispersed. Most of it goes into the ground, and depending on the rigidity of the ball and the ground, this energy may cause the ball to bounce. Some of the energy may appear in the form of sound, produced as the ball hits bottom, and some will manifest as heat. The total energy, however, will not be lost: it will simply have changed form.

REST ENERGY. The values for mechanical energy in the above illustration would most likely be very small; on the other hand, the rest or mass energy of the baseball would be staggering. Given the weight of 0.333 pounds for a regulation baseball, which on Earth converts to 0.15 kg in mass, it would possess enough energy by virtue of its mass to provide a year's worth of electrical power to more than 150,000 American homes. This leads to two obvious questions: how can a mere baseball possess all that energy? And if it does, how can the energy be extracted and put to use?

The answer to the second question is, "By accelerating it to something close to the speed of light"—which is more than 27,000 times faster than the fastest speed ever achieved by humans. (The astronauts on *Apollo 10* in May 1969 reached nearly 25,000 MPH (40,000 km/h), which is more than 33 times the speed of sound but still insignificant when compared to the speed of light.) The answer to the first question lies in the most well-known physics formula of all time: $E = mc^2$

In 1905, Albert Einstein (1879-1955) published his Special Theory of Relativity, which he followed a decade later with his General Theory of Relativity. These works introduced the world to the above-mentioned formula, which holds that energy is equal to mass multiplied by the squared speed of light. This formula gained its widespread prominence due to the many implications of Einstein's Relativity, which quite literally changed humanity's perspective on the universe. Most concrete among those implications was the atom bomb, made possible by the understanding of mass and energy achieved by Einstein.

In fact, $E = mc^2$ is the formula for rest energy, sometimes called mass energy. Though rest energy is "outside" of kinetic and potential energy in the sense that it is not defined by the above-described interactions within the larger system of



As Surya Bonaly goes into a spin on the ice, she draws in her arms and leg, reducing the moment of inertia. Because of the conservation of angular momentum, her angular velocity will increase, meaning that she will spin much faster. *(Bolemian Nomad Picturemakers/Corbis. Reproduced by permission.)*

mechanical energy, its relation to the other forms can be easily shown. All three are defined in terms of mass. Potential energy is equal to mgh, where m is mass, g is gravity, and h is height. Kinetic energy is equal to $\frac{1}{2} mv^2$, where v is velocity. In fact—using a series of steps that will not be demonstrated here—it is possible to directly relate the kinetic and rest energy formulae.

The kinetic energy formula describes the behavior of objects at speeds well below the speed of light, which is 186,000 mi (297,600 km) per second. But at speeds close to that of the speed of light, $\frac{1}{2} mv^2$ does not accurately reflect the energy possessed by the object. For instance, if v were equal to $0.999c$ (where c represents the speed of light), then the application of the formula $\frac{1}{2} mv^2$ would yield a value equal to less than 3% of the object's real energy. In order to calculate the true energy of an object at $0.999c$, it would be necessary to apply a different formula for total energy, one that takes into account the fact that, at such a speed, mass itself becomes energy.

CONSERVATION OF MASS. Mass itself is relative at speeds approaching *c*, and, in fact, becomes greater and greater the closer an object comes to the speed of light. This may seem strange in light of the fact that there is, after all, a law stating that mass is conserved. But mass is only conserved at speeds well below c: as an object approaches 186,000 mi (297,600 km) per second, the rules are altered.

The conservation of mass states that total mass is constant, and is unaffected by factors such as position, velocity, or temperature, in any system that does not exchange any matter with its environment. Yet, at speeds close to *c,* the mass of an object increases dramatically.

In such a situation, the mass would be equal to the rest, or starting mass, of the object divided by $\sqrt{(1 - (v^2/c^2)}$, where v is the object's speed of relative motion. The denominator of this equation will always be less than one, and the greater the value of v, the smaller the value of the denominator. This means that at a speed of *c,* the denominator is zero—in other words, the object's mass is infinite! Obviously, this is not possible, and indeed, what the formula actually shows is that no object can travel faster than the speed of light.

Of particular interest to the present discussion, however, is the fact, established by relativity theory, that mass can be converted into energy. Hence, as noted earlier, a baseball or indeed any object can be converted into energy—and since the formula for rest energy requires that the mass be multiplied by c^2, clearly, even an object of virtually negligible mass can generate a staggering amount of energy. This conversion of mass to energy happens well below the speed of light, in a very small way, when a stick of dynamite explodes. A portion of that stick becomes energy, and the fact that this portion is equal to just 6 parts out of 100 billion indicates the vast proportions of energy available from converted mass.

Other Conservation Laws

In addition to the conservation of energy, as well as the limited conservation of mass, there are laws governing the conservation of momentum, both for an object in linear (straight-line) motion, and for one in angular (rotational) motion. Momentum is a property that a moving body possesses by virtue of its mass and velocity, which determines the amount of force and time required to stop it. Linear momentum is equal to mass multiplied by velocity, and the conservation of linear momentum law states that when the sum of the external force vectors acting on a physical system is equal to zero, the total linear momentum of the system remains unchanged, or conserved.

Angular momentum, or the momentum of an object in rotational motion, is equal to $mr^2\omega$, where m is mass, r is the radius of rotation, and ω (the Greek letter omega) stands for angular velocity. According to the conservation of angular momentum law, when the sum of the external torques acting on a physical system is equal to zero, the total angular momentum of the system remains unchanged. Torque is a force applied around an axis of rotation. When playing the old game of "spin the bottle," for instance, one is applying torque to the bottle and causing it to rotate.

ELECTRIC CHARGE. The conservation of both linear and angular momentum are best explained in the context of real-life examples, provided below. Before going on to those examples, however, it is appropriate here to discuss a conservation law that is outside the realm of everyday experience: the conservation of electric charge, which holds that for an isolated system, the net electric charge is constant.

This law is "outside the realm of everyday experience" such that one cannot experience it through the senses, but at every moment, it is happening everywhere. Every atom has positively charged protons, negatively charged electrons, and uncharged neutrons. Most atoms are neutral, possessing equal numbers of protons and electrons; but, as a result of some disruption, an atom may have more protons than electrons, and thus, become positively charged. Conversely, it may end up with a net negative charge due to a greater number of electrons. But the protons or electrons it released or gained did not simply appear or disappear: they moved from one part of the system to another—that is, from one atom to another atom, or to several other atoms.

Throughout these changes, the charge of each proton and electron remains the same, and the net charge of the system is always the sum of its positive and negative charges. Thus, it is impossible for any electrical charge in the universe to be smaller than that of a proton or electron. Likewise, throughout the universe, there is always the same number of negative and positive electrical charges: just as energy changes form, the charges simply change position.

There are also conservation laws describing the behavior of subatomic particles, such as the positron and the neutrino. However, the most significant of the conservation laws are those involving energy (and mass, though with the limitations discussed above), linear momentum, angular momentum, and electrical charge.

REAL-LIFE APPLICATIONS

Conservation of Linear Momentum: Rifles and Rockets

FIRING A RIFLE. The conservation of linear momentum is reflected in operations as simple as the recoil of a rifle when it is fired, and in those as complex as the propulsion of a rocket through space. In accordance with the conservation of momentum, the momentum of a system must be the same after it undergoes an operation as it was before the process began. Before firing, the momentum of a rifle and bullet is zero, and therefore, the rifle-bullet system must return to that same zero-level of momentum after it is fired. Thus, the momentum of the bullet must be matched—and "cancelled" within the system under study—by a corresponding backward momentum.

When a person shooting a gun pulls the trigger, it releases the bullet, which flies out of the barrel toward the target. The bullet has mass and velocity, and it clearly has momentum; but this is only half of the story. At the same time it is fired, the rifle produces a "kick," or sharp jolt, against the shoulder of the person who fired it. This backward kick, with a velocity in the opposite direction of the bullet's trajectory, has a momentum exactly the same as that of the bullet itself: hence, momentum is conserved.

But how can the rearward kick have the same momentum as that of the bullet? After all, the bullet can kill a person, whereas, if one holds the rifle correctly, the kick will not even cause any injury. The answer lies in several properties of linear momentum. First of all, as noted earlier, momentum is equal to mass multiplied by velocity; the actual proportions of mass and velocity, however, are not important as long as the backward momentum is the same as the forward momentum. The bullet is an object of relatively small mass and high velocity, whereas the rifle is much larger in mass, and hence, its rearward velocity is correspondingly small.

In addition, there is the element of impulse, or change in momentum. Impulse is the product of force multiplied by change or interval in time. Again, the proportions of force and time interval do not matter, as long as they are equal to the momentum change—that is, the difference in momentum that occurs when the rifle is fired. To avoid injury to one's shoulder, clearly force must be minimized, and for this to happen, time interval must be extended.

If one were to fire the rifle with the stock (the rear end of the rifle) held at some distance from one's shoulder, it would kick back and could very well produce a serious injury. This is because the force was delivered over a very short time interval—in other words, force was maximized and time interval minimized. However, if one holds the rifle stock firmly against one's shoulder, this slows down the delivery of the kick, thus maximizing time interval and minimizing force.

ROCKETING THROUGH SPACE. Contrary to popular belief, rockets do not move by pushing against a surface such as a launchpad. If that were the case, then a rocket would have nothing to propel it once it had been launched, and certainly there would be no way for a rocket to move through the vacuum of outer space. Instead, what propels a rocket is the conservation of momentum.

Upon ignition, the rocket sends exhaust gases shooting downward at a high rate of velocity. The gases themselves have mass, and thus, they have momentum. To balance this downward momentum, the rocket moves upward—though, because its mass is greater than that of the gases it expels, it will not move at a velocity as high as that of the gases. Once again, the upward or forward momentum is exactly the same as the downward or backward momentum, and linear momentum is conserved.

Rather than needing something to push against, a rocket in fact performs best in outer space, where there is nothing—neither launchpad nor even air—against which to push. Not only is "pushing" irrelevant to the operation of

KEY TERMS

CONSERVATION LAWS: A set of principles describing physical properties that remain constant—that is, are conserved—throughout the various processes that occur in the physical world. The most significant of these laws concerns the conservation of energy (as well as, with qualifications, the conservation of mass); conservation of linear momentum; conservation of angular momentum; and conservation of electrical charge.

CONSERVATION OF ANGULAR MOMENTUM: A physical law stating that when the sum of the external torques acting on a physical system is equal to zero, the total angular momentum of the system remains unchanged. Angular momentum is the momentum of an object in rotational motion, and torque is a force applied around an axis of rotation.

CONSERVATION OF ELECTRICAL CHARGE: A physical law which holds that for an isolated system, the net electrical charge is constant.

CONSERVATION OF ENERGY: A law of physics stating that within a system isolated from all other outside factors, the total amount of energy remains the same, though transformations of energy from one form to another take place.

CONSERVATION OF LINEAR MOMENTUM: A physical law stating that when the sum of the external force vectors acting on a physical system is equal to zero, the total linear momentum of the system remains unchanged—or is conserved.

CONSERVATION OF MASS: A physical principle stating that total mass is constant, and is unaffected by factors such as position, velocity, or temperature, in any system that does not exchange any matter with its environment. Unlike the other conservation laws, however, conservation of mass is not universally applicable, but applies only at speeds significant lower than that of light—186,000 mi (297,600 km) per second. Close to the speed of light, mass begins converting to energy.

CONSERVE: In physics, "to conserve" something means "to result in no net loss of" that particular component. It is possible that within a given system, the component may change form or position, but as long as the net value of the component remains the same, it has been conserved.

FRICTION: The force that resists motion when the surface of one object comes into contact with the surface of another.

MOMENTUM: A property that a moving body possesses by virtue of its mass and velocity, which determines the amount of force and time required to stop it.

SYSTEM: In physics, the term "system" usually refers to any set of physical interactions isolated from the rest of the universe. Anything outside of the system, including all factors and forces irrelevant to a discussion of that system, is known as the environment.

the rocket, but the rocket moves much more efficiently without the presence of air resistance. In the same way, on the relatively frictionless surface of an ice-skating rink, conservation of linear momentum (and hence, the process that makes possible the flight of a rocket through space) is easy to demonstrate.

If, while standing on the ice, one throws an object in one direction, one will be pushed in the opposite direction with a corresponding level of momentum. However, since a person's mass is presumably greater than that of the object thrown, the rearward velocity (and, therefore, distance) will be smaller.

Friction, as noted earlier, is not the only force that counters conservation of linear momentum on Earth: so too does gravity, and thus, once again, a rocket operates much better in space than it does when under the influence of Earth's gravitational field. If a bullet is fired at a bottle thrown into the air, the linear momentum of the spent bullet and the shattered pieces of glass in the infinitesimal moment just after the collision will be the same as that of the bullet and the bottle a moment before impact. An instant later, however, gravity will accelerate the bullet and the pieces downward, thus leading to a change in total momentum.

Conservation of Angular Momentum: Skaters and Other Spinners

As noted earlier, angular momentum is equal to $mr^2\omega$, where m is mass, r is the radius of rotation, and ω stands for angular velocity. In fact, the first two quantities, mr^2, are together known as moment of inertia. For an object in rotation, moment of inertia is the property whereby objects further from the axis of rotation move faster, and thus, contribute a greater share to the overall kinetic energy of the body.

One of the most oft-cited examples of angular momentum—and of its conservation—involves a skater or ballet dancer executing a spin. As the skater begins the spin, she has one leg planted on the ice, with the other stretched behind her. Likewise, her arms are outstretched, thus creating a large moment of inertia. But when she goes into the spin, she draws in her arms and leg, reducing the moment of inertia. In accordance with conservation of angular momentum, $mr^2\omega$ will remain constant, and therefore, her angular velocity will increase, meaning that she will spin much faster.

CONSTANT ORIENTATION. The motion of a spinning top and a Frisbee in flight also illustrate the conservation of angular momentum. Particularly interesting is the tendency of such an object to maintain a constant orientation. Thus, a top remains perfectly vertical while it spins, and only loses its orientation once friction from the floor dissipates its velocity and brings it to a stop. On a frictionless surface, however, it would remain spinning—and therefore upright—forever.

A Frisbee thrown without spin does not provide much entertainment; it will simply fall to the ground like any other object. But if it is tossed with the proper spin, delivered from the wrist, conservation of angular momentum will keep it in a horizontal position as it flies through the air. Once again, the Frisbee will eventually be brought to ground by the forces of air resistance and gravity, but a Frisbee hurled through empty space would keep spinning for eternity.

WHERE TO LEARN MORE

Beiser, Arthur. *Physics,* 5th ed. Reading, MA: Addison-Wesley, 1991.

"Conservation Laws: An Online Physics Textbook" (Web site). <http://www.lightandmatter.com/area1book2.html> (March 12, 2001).

"Conservation Laws: The Most Powerful Laws of Physics" (Web site). <http://webug.physics.uiuc.edu/courses/phys150/fall99/slides/lect0 7/> (March 12, 2001).

"Conservation of Energy." NASA (Web site). <http://www.grc.nasa.gov/WWW/K-12/airplane/thermo1f.html> (March 12, 2001).

Elkana, Yehuda. *The Discovery of the Conservation of Energy.* With a foreword by I. Bernard Cohen. Cambridge, MA: Harvard University Press, 1974.

"Momentum and Its Conservation" (Web site). <http://www.glenbrook.k12.il.us/gbssci/phys/Class/momentum/momtoc. html> (March 12, 2001).

Rutherford, F. James; Gerald Holton; and Fletcher G. Watson. *Project Physics.* New York: Holt, Rinehart, and Winston, 1981.

Suplee, Curt. *Everyday Science Explained.* Washington, D.C.: National Geographic Society, 1996.

KINEMATICS AND PARTICLE DYNAMICS

MOMENTUM

CENTRIPETAL FORCE

FRICTION

LAWS OF MOTION

GRAVITY

PROJECTILE MOTION

TORQUE

MOMENTUM

CONCEPT

The faster an object is moving—whether it be a baseball, an automobile, or a particle of matter—the harder it is to stop. This is a reflection of momentum, or specifically, linear momentum, which is equal to mass multiplied by velocity. Like other aspects of matter and motion, momentum is conserved, meaning that when the vector sum of outside forces equals zero, no net linear momentum within a system is ever lost or gained. A third important concept is impulse, the product of force multiplied by length in time. Impulse, also defined as a change in momentum, is reflected in the proper methods for hitting a baseball with force or surviving a car crash.

HOW IT WORKS

Like many other aspects of physics, the word "momentum" is a part of everyday life. The common meaning of momentum, however, unlike many other physics terms, is relatively consistent with its scientific meaning. In terms of formula, momentum is equal to the product of mass and velocity, and the greater the value of that product, the greater the momentum.

Consider the term "momentum" outside the world of physics, as applied, for example, in the realm of politics. If a presidential candidate sees a gain in public-opinion polls, then wins a debate and embarks on a whirlwind speaking tour, the media comments that he has "gained momentum." As with momentum in the framework of physics, what these commentators mean is that the candidate will be hard to stop—or to carry the analogy further, that he is doing enough of the right things (thus gaining "mass"), and doing them quickly enough, thereby gaining velocity.

Momentum and Inertia

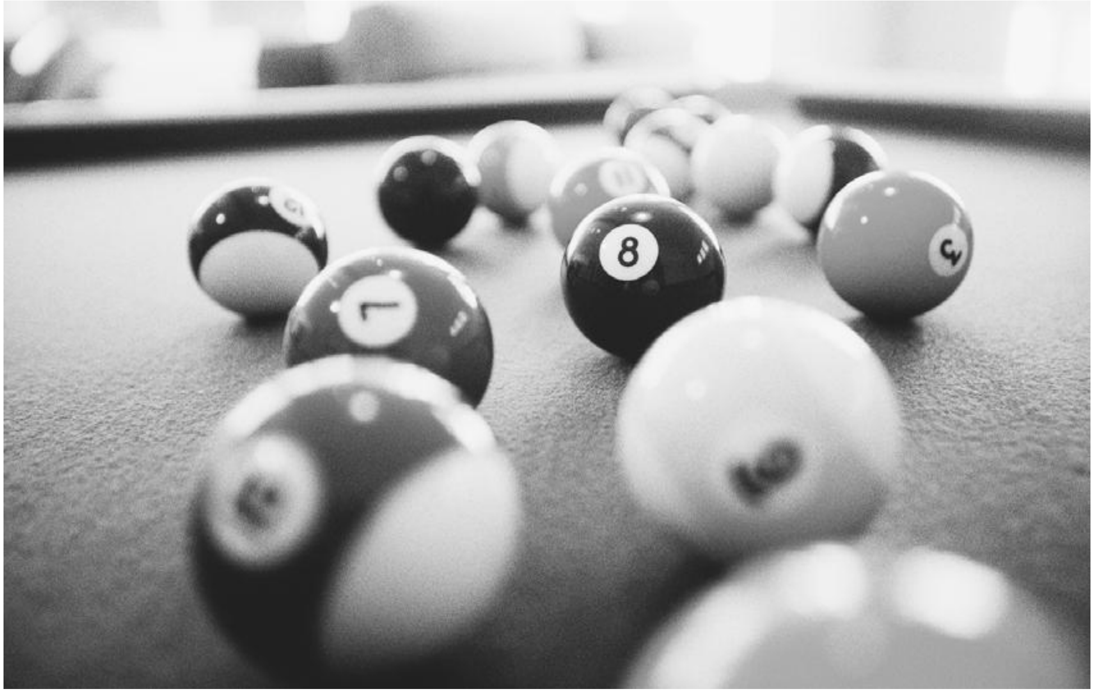
It might be tempting to confuse momentum with another physical concept, inertia. Inertia, as defined by the second law of motion, is the tendency of an object in motion to remain in motion, and of an object at rest to remain at rest. Momentum, by definition, involves a body in motion, and can be defined as the tendency of a body in motion to continue moving at a constant velocity.

Not only does momentum differ from inertia in that it relates exclusively to objects in motion, but (as will be discussed below) the component of velocity in the formula for momentum makes it a vector—that is, a quantity that possesses both magnitude and direction. There is at least one factor that momentum very clearly has in common with inertia: mass, a measure of inertia indicating the resistance of an object to a change in its motion.

Mass and Weight

Unlike velocity, mass is a scalar, a quantity that possesses magnitude without direction. Mass is often confused with weight, a vector quantity equal to its mass multiplied by the downward acceleration due to gravity. The weight of an object changes according to the gravitational force of the planet or other celestial body on which it is measured. Hence, the mass of a person on the Moon would be the same as it is on Earth, whereas the person's weight would be considerably less, due to the smaller gravitational pull of the Moon.

Given the unchanging quality of mass as opposed to weight, as well as the fact that scientists themselves prefer the much simpler metric



When billiard balls collide, their hardness results in an elastic collision—one in which kinetic energy is conserved. *(Photograph by John-Marshall Mantel/Corbis. Reproduced by permission.)*

system, metric units will generally be used in the following discussion. Where warranted, of course, conversion to English or British units (for example, the pound, a unit of weight) will be provided. However, since the English unit of mass, the slug, is even more unfamiliar to most Americans than its metric equivalent, the kilogram, there is little point in converting kilos into slugs.

Velocity and Speed

Not only is momentum often confused with inertia, and mass with weight, but in the everyday world the concepts of velocity and speed tend to be blurred. Speed is the rate at which the position of an object changes over a given period of time, expressed in terms such as "50 MPH." It is a scalar quantity.

Velocity, by contrast, is a vector. If one were to say "50 miles per hour toward the northeast," this would be an expression of velocity. Vectors are typically designated in bold, without italics; thus velocity is typically abbreviated **v**. Scalars, on the other hand, are rendered in italics. Hence, the formula for momentum is usually shown as *m***v**.

Linear Momentum and Its Conservation

Momentum itself is sometimes designated as *p*. It should be stressed that the form of momentum discussed here is strictly linear, or straight-line, momentum, in contrast to angular momentum, more properly discussed within the framework of rotational motion.

Both angular and linear momentum abide by what are known as conservation laws. These are statements concerning quantities that, under certain conditions, remain constant or unchanging. The conservation of linear momentum law states that when the sum of the external force vectors acting on a physical system is equal to zero, the total linear momentum of the system remains unchanged—or conserved.

The conservation of linear momentum is reflected both in the recoil of a rifle and in the propulsion of a rocket through space. When a rifle is fired, it produces a "kick"—that is, a sharp jolt to the shoulder of the person who has fired it—corresponding to the momentum of the bullet. Why, then, does the "kick" not knock a person's shoulder off the way a bullet would? Because the rifle's mass is much greater than that of the bullet, meaning that its velocity is much smaller.

As for rockets, they do not—contrary to popular belief—move by pushing against a surface, such as a launch pad. If that were the case, then a rocket would have nothing to propel it once it is launched, and certainly there would be no way for a rocket to move through the vacuum of outer space. Instead, as it burns fuel, the rocket expels exhaust gases that exert a backward momentum, and the rocket itself travels forward with a corresponding degree of momentum.

SYSTEMS

Here, "system" refers to any set of physical interactions isolated from the rest of the universe. Anything outside of the system, including all factors and forces irrelevant to a discussion of that system, is known as the environment. In the pool-table illustration shown earlier, the interaction of the billiard balls in terms of momentum is the system under discussion.

It is possible to reduce a system even further for purposes of clarity: hence, one might specify that the system consists only of the pool balls, the force applied to them, and the resulting momentum interactions. Thus, we will ignore the friction of the pool table's surface, and the assumption will be that the balls are rolling across a frictionless plane.

IMPULSE

For an object to have momentum, some force must have set it in motion, and that force must have been applied over a period of time. Likewise, when the object experiences a collision or any other event that changes its momentum, that change may be described in terms of a certain amount of force applied over a certain period of time. Force multiplied by time interval is impulse, expressed in the formula $F \cdot \delta t$, where **F** is force, δ (the Greek letter delta) means "a change" or "change in..."; and *t* is time.

As with momentum itself, impulse is a vector quantity. Whereas the vector component of momentum is velocity, the vector quantity in impulse is force. The force component of impulse can be used to derive the relationship between impulse and change in momentum. According to the second law of motion, $\mathbf{F} = m\,\mathbf{a}$; that is, force is equal to mass multiplied by acceleration. Acceleration can be defined as a change

WHEN PARACHUTISTS LAND, THEY KEEP THEIR KNEES BENT AND OFTEN ROLL OVER—ALL IN AN EFFORT TO LENGTHEN THE PERIOD OF THE FORCE OF IMPACT, THUS REDUCING ITS EFFECTS. *(Photograph by James A. Sugar/Corbis. Reproduced by permission.)*

in velocity over a change or interval in time. Expressed as a formula, this is

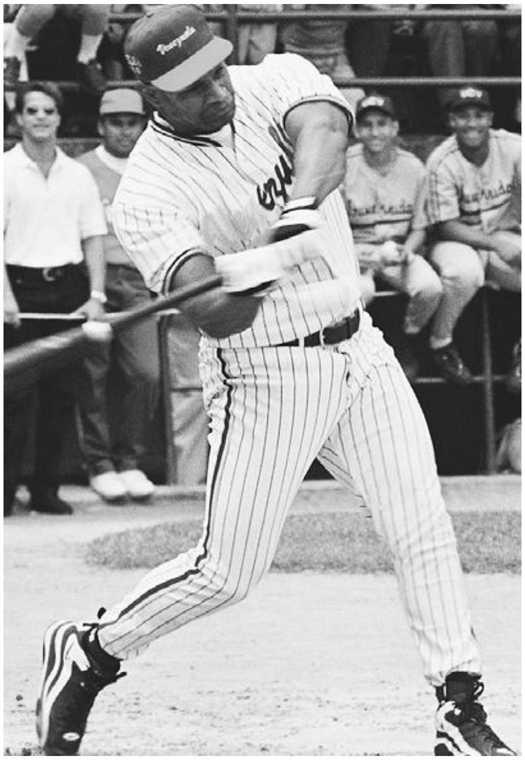
$$a = \frac{\Delta v}{\Delta t}.$$

Thus, force is equal to

$$m\left(\frac{\Delta v}{\Delta t}\right),$$

an equation that can be rewritten as $\mathbf{F}\delta t = m\delta\mathbf{v}$. In other words, impulse is equal to change in momentum.

This relationship between impulse and momentum change, derived here in mathematical terms, will be discussed below in light of several well-known examples from the real world. Note that the metric units of impulse and momentum are actually interchangeable, though they are typically expressed in different forms, for the purpose of convenience. Hence, momentum is usually rendered in terms of kilogram-meters-per-second (kg • m/s), whereas impulse is typically shown as newton-seconds (N • s). In the English system, momentum is shown in units of slug-feet per-



As Sammy Sosa's bat hits this ball, it applies a tremendous momentum change to the ball. After contact with the ball, Sosa will continue his swing, thereby contributing to the momentum change and allowing the ball to travel farther. *(AFP/Corbis. Reproduced by permission.)*

second, and impulse in terms of the pound-second.

REAL-LIFE APPLICATIONS

When Two Objects Collide

Two moving objects, both possessing momentum by virtue of their mass and velocity, collide with one another. Within the system created by their collision, there is a total momentum $M\mathbf{V}$ that is equal to their combined mass and the vector sum of their velocity.

This is the case with any system: the total momentum is the sum of the various individual momentum products. In terms of a formula, this is expressed as $M\mathbf{V} = m_1\mathbf{v}_1 + m_2\mathbf{v}_2 + m_3\mathbf{v}_3 + ...$ and so on. As noted earlier, the total momentum will be conserved; however, the actual distribution of momentum within the system may change.

TWO LUMPS OF CLAY. Consider the behavior of two lumps of clay, thrown at one another so that they collide head-on. Due to the properties of clay as a substance, the two lumps will tend to stick. Assuming the lumps are not of equal mass, they will continue traveling in the same direction as the lump with greater momentum.

As they meet, the two lumps form a larger mass $M\mathbf{V}$ that is equal to the sum of their two individual masses. Once again, $M\mathbf{V} = m_1\mathbf{v}_1 + m_2\mathbf{v}_2$. The M in $M\mathbf{V}$ is the sum of the smaller values m, and the $\mathbf{V}$ is the *vector sum* of velocity. Whereas M is larger than m_1 or m_2—the reason being that scalars are simply added like ordinary numbers—V is smaller than $\mathbf{v}_1$ or $\mathbf{v}_2$. This lower number for net velocity as compared to particle velocity will always occur when two objects are moving in opposite directions. (If the objects are moving in the same direction, V will have a value between that of $\mathbf{v}_1$ and $\mathbf{v}_2$.)

To add the vector sum of the two lumps in collision, it is best to make a diagram showing the bodies moving toward one another, with arrows illustrating the direction of velocity. By convention, in such diagrams the velocity of an object moving to the right is rendered as a positive number, and that of an object moving to the left is shown with a negative number. It is therefore easier to interpret the results if the object with the larger momentum is shown moving to the right.

The value of $\mathbf{V}$ will move in the same direction as the lump with greater momentum. But since the two lumps are moving in opposite directions, the momentum of the smaller lump will cancel out a portion of the greater lump's momentum—much as a negative number, when added to a positive number of greater magnitude, cancels out part of the positive number's value. They will continue traveling in the direction of the lump with greater momentum, now with a combined mass equal to the arithmetic sum of their masses, but with a velocity much smaller than either had before impact.

BILLIARD BALLS. The game of pool provides an example of a collision in which one object, the cue ball, is moving, while the other—known as the object ball—is stationary. Due to the hardness of pool balls, and their tendency not to stick to one another, this is also an example of an almost perfectly elastic collision—one in which kinetic energy is conserved.

The colliding lumps of clay, on the other hand, are an excellent example of an inelastic collision, or one in which kinetic energy is not conserved. The total energy in a given system, such as that created by the two lumps of clay in collision, is conserved; however, kinetic energy may be transformed, for instance, into heat energy and/or sound energy as a result of collision. Whereas inelastic collisions involve soft, sticky objects, elastic collisions involve rigid, non-sticky objects.

Kinetic energy and momentum both involve components of velocity and mass: **p** (momentum) is equal to $m\mathbf{v}$, and KE (kinetic energy) equals ½ $m\mathbf{v}^2$. Due to the elastic nature of pool-ball collisions, when the cue ball strikes the object ball, it transfers its velocity to the latter. Their masses are the same, and therefore the resulting momentum and kinetic energy of the object ball will be the same as that possessed by the cue ball prior to impact.

If the cue ball has transferred all of its velocity to the object ball, does that mean it has stopped moving? It does. Assuming that the interaction between the cue ball and the object ball constitutes a closed system, there is no other source from which the cue ball can acquire velocity, so its velocity must be zero.

It should be noted that this illustration treats pool-ball collisions as though they were 100% elastic, though in fact, a portion of kinetic energy in these collisions is transformed into heat and sound. Also, for a cue ball to transfer all of its velocity to the object ball, it must hit it straight-on. If the balls hit off-center, not only will the object ball move after impact, but the cue ball will continue to move—roughly at 90° to a line drawn through the centers of the two balls at the moment of impact.

Impulse: Breaking or Building the Impact

When a cue ball hits an object ball in pool, it is safe to assume that a powerful impact is desired. The same is true of a bat hitting a baseball. But what about situations in which a powerful impact is not desired—as for instance when cars are crashing? There is, in fact, a relationship between impulse, momentum change, transfer of kinetic energy, and the impact—desirable or undesirable—experienced as a result.

Impulse, again, is equal to momentum change—and also equal to force multiplied by time interval (or change in time). This means that the greater the force and the greater the amount of time over which it is applied, the greater the momentum change. Even more interesting is the fact that one can achieve the same momentum change with differing levels of force and time interval. In other words, a relatively low degree of force applied over a relatively long period of time would produce the same momentum change as a relatively high amount of force over a relatively short period of time.

The conservation of kinetic energy in a collision is, as noted earlier, a function of the relative elasticity of that collision. The question of whether KE is transferred has nothing to do with impulse. On the other hand, the question of how KE is transferred—or, even more specifically, the interval over which the transfer takes place—is very much related to impulse.

Kinetic energy, again, is equal to ½ $m\mathbf{v}^2$. If a moving car were to hit a stationary car head-on, it would transfer a quantity of kinetic energy to the stationary car equal to one-half its own mass multiplied by the square of its velocity. (This, of course, assumes that the collision is perfectly elastic, and that the mass of the cars is exactly equal.) A transfer of KE would also occur if two moving cars hit one another head-on, especially in a highly elastic collision. Assuming one car had considerably greater mass and velocity than the other, a high degree of kinetic energy would be transferred—which could have deadly consequences for the people in the car with less mass and velocity. Even with cars of equal mass, however, a high rate of acceleration can bring about a potentially lethal degree of force.

CRUMPLE ZONES IN CARS. In a highly elastic car crash, two automobiles would bounce or rebound off one another. This would mean a dramatic change in direction—a reversal, in fact—hence, a sudden change in velocity and therefore momentum. In other words, the figure for $m\delta\mathrm{v}$ would be high, and so would that for impulse, $\mathrm{F}\delta\mathrm{t}$.

On the other hand, it is possible to have a highly inelastic car crash, accompanied by a small change in momentum. It may seem logical to think that, in a crash situation, it would be better for two cars to bounce off one another than

KEY TERMS

ACCELERATION: A change velocity. Acceleration can be expressed as a formula $\delta v/\delta t$—that is, change in velocity divided by change, or interval, in time.

CONSERVATION OF LINEAR MOMENTUM: A physical law, which states that when the sum of the external force vectors acting on a physical system is equal to zero, the total linear momentum of the system remains unchanged—or is conserved.

CONSERVE: In physics, "to conserve" something (for example, momentum or kinetic energy) means "to result in no net loss of" that particular component. It is possible that within a given system, one type of energy may be transformed into another type, but the net energy in the system will remain the same.

ELASTIC COLLISION: A collision in which kinetic energy is conserved. Typically elastic collisions involve rigid, non-sticky objects such as pool balls. At the other extreme is an inelastic collision.

IMPULSE: The amount of force and time required to cause a change in momentum. Impulse is the product of force multiplied by a change, or interval, in time ($\mathbf{F}\delta t$): the greater the momentum, the greater the force needed to change it, and the longer the period of time over which it must be applied.

INELASTIC COLLISION: A collision in which kinetic energy is not conserved. (The total energy is conserved: kinetic energy itself, however, may be transformed into heat energy or sound energy.) Typically, inelastic collisions involve non-rigid, sticky objects—for instance, lumps of clay. At the other extreme is an elastic collision.

INERTIA: The tendency of an object in motion to remain in motion, and of an object at rest to remain at rest.

KINETIC ENERGY: The energy an object possesses by virtue of its motion.

MASS: A measure of inertia, indicating the resistance of an object to a change in its

for them to crumple together. In fact, however, the latter option is preferable. When the cars crumple rather than rebounding, they do not experience a reversal in direction. They do experience a change in speed, of course, but the momentum change is far less than it would be if they rebounded.

Furthermore, crumpling lengthens the amount of time during which the change in velocity occurs, and thus reduces impulse. But even with the reduced impulse of this momentum change, it is possible to further reduce the effect of force, another aspect of impact. Remember that $m\delta\mathbf{v} = \mathbf{F}\delta t$: the value of force and time interval do not matter, as long as their product is equal to the momentum change. Because **F** and δt are inversely proportional, an increase in impact time will reduce the effects of force.

For this reason, car manufacturers actually design and build into their cars a feature known as a crumple zone. A crumple zone—and there are usually several in a single automobile—is a section in which the materials are put together in such a way as to ensure that they will crumple when the car experiences a collision. Of course, the entire car cannot be one big crumple zone—this would be fatal for the driver and riders; however, the incorporation of crumple zones at key points can greatly reduce the effect of the force a car and its occupants must endure in a crash.

Another major reason for crumple zones is to keep the passenger compartment of the car

KEY TERMS CONTINUED

motion—including a change in velocity. A kilogram is a unit of mass, whereas a pound is a unit of weight.

MOMENTUM: A property that a moving body possesses by virtue of its mass and velocity, which determines the amount of force and time (impulse) required to stop it. Momentum—actually linear momentum, as opposed to the angular momentum of an object in rotational motion—is equal to mass multiplied by velocity.

SCALAR: A quantity that possesses only magnitude, with no specific direction—as contrasted with a vector, which possesses both magnitude and direction. Scalar quantities are usually expressed in italicized letters, thus: *m* (mass).

SPEED: The rate at which the position of an object changes over a given period of time.

SYSTEM: In physics, the term "system" usually refers to any set of physical interactions isolated from the rest of the universe. Anything outside of the system, including all factors and forces irrelevant to a discussion of that system, is known as the environment.

VECTOR: A quantity that possesses both magnitude and direction—as contrasted with a scalar, which possesses magnitude without direction. Vector quantities are usually expressed in bold, non-italicized letters, thus: **F** (force). They may also be shown by placing an arrow over the letter designating the specific property, as for instance v for velocity.

VECTOR SUM: A calculation that yields the net result of all the vectors applied in a particular situation. In the case of momentum, the vector component is velocity. The best method is to make a diagram showing bodies in collision, with arrows illustrating the direction of velocity. On such a diagram, motion to the right is assigned a positive value, and to the left a negative value.

VELOCITY: The speed of an object in a particular direction.

intact. Many injuries are caused when the body of the car intrudes on the space of the occupants—as, for instance, when the floor buckles, or when the dashboard is pushed deep into the passenger compartment. Obviously, it is preferable to avoid this by allowing the fender to collapse.

REDUCING IMPULSE: SAVING LIVES, BONES, AND WATER BALLOONS. An airbag is another way of minimizing force in a car accident, in this case by reducing the time over which the occupants move forward toward the dashboard or windshield. The airbag rapidly inflates, and just as rapidly begins to deflate, within the split-second that separates the car's collision and a person's collision with part of the car. As it deflates, it is receding toward the dashboard even as the driver's or passenger's body is being hurled toward the dashboard. It slows down impact, extending the amount of time during which the force is distributed.

By the same token, a skydiver or paratrooper does not hit the ground with legs outstretched: he or she would be likely to suffer a broken bone or worse from such a foolish stunt. Rather, as a parachutist prepares to land, he or she keeps knees bent, and upon impact immediately rolls over to the side. Thus, instead of experiencing the force of impact over a short period of time, the parachutist lengthens the amount of time that force is experienced, which reduces its effects.

The same principle applies if one were catching a water balloon. In order to keep it from bursting, one needs to catch the balloon in midair, then bring it to a stop slowly by "traveling" with it for a few feet before reducing its momentum down to zero. Once again, there is no way around the fact that one is attempting to bring about a substantial momentum change—a change equal in value to the momentum of the object in movement. Nonetheless, by increasing the time component of impulse, one reduces the effects of force.

In old *Superman* comics, the "Man of Steel" often caught unfortunate people who had fallen, or been pushed, out of tall buildings. The cartoons usually showed him, at a stationary position in midair, catching the person before he or she could hit the ground. In fact, this would not save their lives: the force component of the sudden momentum change involved in being caught would be enough to kill the person. Of course, it is a bit absurd to quibble over scientific accuracy in *Superman,* but in order to make the situation more plausible, the "Man of Steel" should have been shown catching the person, then slowly following through on the trajectory of the fall toward earth.

THE CRACK OF THE BAT: INCREASING IMPULSE. But what if—to once again turn the tables—a strong force is desired? This time, rather than two pool balls striking one another, consider what happens when a batter hits a baseball. Once more, the correlation between momentum change and impulse can create an advantage, if used properly.

As the pitcher hurls the ball toward home plate, it has a certain momentum; indeed, a pitch thrown by a major-league player can send the ball toward the batter at speeds around 100 MPH (160 km/h)—a ball having considerable momentum). In order to hit a line drive or "knock the ball out of the park," the batter must therefore cause a significant change in momentum.

Consider the momentum change in terms of the impulse components. The batter can only apply so much force, but it is possible to magnify impulse greatly by increasing the amount of time over which the force is delivered. This is known in sports—and it applies as much in tennis or golf as in baseball—as "following through." By increasing the time of impact, the batter has increased impulse and thus, momentum change. Obviously, the mass of the ball has not been altered; the difference, then, is a change in velocity.

How is it possible that in earlier examples, the effects of force were decreased by increasing the time interval, whereas in the baseball illustration, an increase in time interval resulted in a more powerful impact? The answer relates to differences in direction and elasticity. The baseball and the bat are colliding head-on in a relatively elastic situation; by contrast, crumpling cars are inelastic. In the example of a person catching a water balloon, the catcher is moving in the same direction as the balloon, thus reducing momentum change. Even in the case of the paratrooper, the ground is stationary; it does not move toward the parachutist in the way that the baseball moves toward the bat.

WHERE TO LEARN MORE

Beiser, Arthur. *Physics,* 5th ed. Reading, MA: Addison-Wesley, 1991.

Bonnet, Robert L. and Dan Keen. *Science Fair Projects: Physics.* Illustrated by Frances Zweifel. New York: Sterling, 1999.

Fleisher, Paul. *Objects in Motion: Principles of Classical Mechanics.* Minneapolis, MN: Lerner Publications, 2002.

Gardner, Robert. *Experimenting with Science in Sports.* New York: F. Watts, 1993.

"Lesson 1: The Impulse Momentum Change Theorem" (Web site) <http://www.glenbrook.k12.il.us/gbssci/phys/Class/momentum/u411a.html> (March 19, 2001).

"Momentum" (Web site). <http://id.mind.net/~zona/mstm/physics/mechanics/momentum/momentum.html> (March 19, 2001).

Physlink.com (Web site). <http://www.physlink.com> (March 7, 2001).

Rutherford, F. James; Gerald Holton; and Fletcher G. Watson. *Project Physics.* New York: Holt, Rinehart, and Winston, 1981.

Schrier, Eric and William F. Allman. *Newton at the Bat: The Science in Sports.* New York: Charles Scribner's Sons, 1984.

Zubrowski, Bernie. *Raceways: Having Fun with Balls and Tracks.* Illustrated by Roy Doty. New York: William Morrow, 1985.

CENTRIPETAL FORCE

CONCEPT

Most people have heard of centripetal and centrifugal force. Though it may be somewhat difficult to keep track of which is which, chances are anyone who has heard of the two concepts remembers that one is the tendency of objects in rotation to move inward, and the other is the tendency of rotating objects to move outward. It may come as a surprise, then, to learn that there is no such thing, strictly speaking, as centrifugal (outward) force. There is only centripetal (inward) force and the inertia that makes objects in rotation under certain situations move outward, for example, a car making a turn, the movement of a roller coaster— even the spinning of a centrifuge.

HOW IT WORKS

Like many other principles in physics, centripetal force ultimately goes back to a few simple precepts relating to the basics of motion. Consider an object in uniform circular motion: an object moves around the center of a circle so that its speed is constant or unchanging.

The formula for speed—or rather, average speed—is distance divided by time; hence, people say, for instance, "miles (or kilometers) per hour." In the case of an object making a circle, distance is equal to the circumference, or distance around, the circle. From geometry, we know that the formula for calculating the circumference of a circle is $2\pi r$, where r is the radius, or the distance from the circumference to the center. The figure π may be rendered as 3.141592..., though in fact, it is an irrational number: the decimal figures continue forever without repetition or pattern.

From the above, it can be discerned that the formula for the average speed of an object moving around a circle is $2\pi r$ divided by time. Furthermore, we can see that there is a proportional relationship between radius and average speed. If the radius of a circle is doubled, but an object at the circle's periphery makes one complete revolution in the same amount of time as before, this means that the average speed has doubled as well. This can be shown by setting up two circles, one with a radius of 2, the other with a radius of 4, and using some arbitrary period of time—say, 2 seconds.

The above conclusion carries with it an interesting implication with regard to speeds at different points along the radius of a circle. Rather than comparing two points moving around the circumferences of two different circles—one twice as big as the other—in the same period of time, these two points could be on the same circle: one at the periphery, and one exactly halfway along the radius. Assuming they both traveled a complete circle in the same period of time, the proportional relationship described earlier would apply. This means, then, that the further out on the circle one goes, the greater the average speed.

VELOCITY = SPEED + DIRECTION

Speed is a scalar, meaning that it has magnitude but no specific direction; by contrast, velocity is a vector—a quantity with both a magnitude (that is, speed) and a direction. For an object in circular motion, the direction of velocity is the same as that in which the object is moving at any given point. Consider the example of the city of Atlanta, Georgia, and Interstate-285, one of several instances in which a city is surrounded by a "loop" highway. Local traffic reporters avoid giv-

Typically, a centrifuge consists of a base; a rotating tube perpendicular to the base; and vials attached by movable centrifuge arms to the rotating tube. The movable arms are hinged at the top of the rotating tube, and thus can move upward at an angle approaching 90° to the tube. When the tube begins to spin, centripetal force pulls the material in the vials toward the center. *(Photograph by Charles D. Winters. National Audubon Society Collection/Photo Researchers, Inc. Reproduced by permission.)*

ing mere directional coordinates for spots on that highway (for instance, "southbound on 285"), because the area where traffic moves south depends on whether one is moving clockwise or counterclockwise. Hence, reporters usually say "southbound on the outer loop."

As with cars on I-285, the direction of the velocity vector for an object moving around a circle is a function entirely of its position and the direction of movement—clockwise or counterclockwise—for the circle itself. The direction of the individual velocity vector at any given point may be described as tangential; that is, describing a tangent, or a line that touches the circle at just one point. (By definition, a tangent line cannot intersect the circle.)

It follows, then, that the direction of an object in movement around a circle is changing; hence, its velocity is also changing—and this in turn means that it is experiencing acceleration. As with the subject of centripetal force and "centrifugal force," most people have a mistaken view of acceleration, believing that it refers only to an increase in speed. In fact, acceleration is a change in velocity, and can thus refer either to a change in speed or direction. Nor must that change be a positive one; in other words, an object undergoing a reduction in speed is also experiencing acceleration.

The acceleration of an object in rotational motion is always toward the center of the circle. This may appear to go against common sense, which should indicate that acceleration moves in the same direction as velocity, but it can, in fact, be proven in a number of ways. One method would be by the addition of vectors, but a "hands-on" demonstration may be more enlightening than an abstract geometrical proof.

It is possible to make a simple accelerometer, a device for measuring acceleration, with a lit candle inside a glass. The candle should be standing at a 90°-angle to the bottom of the glass, attached to it by hot wax as you would affix a burning candle to a plate. When you hold the candle level, the flame points upward; but if you spin the glass in a circle, the flame will point toward the center of that circle—in the direction of acceleration.

Mass × Acceleration = Force

Since we have shown that acceleration exists for an object spinning around a circle, it is then possible for us to prove that the object experiences some type of force. The proof for this assertion lies in the second law of motion, which defines force as the product of mass and acceleration: hence, where there is acceleration and mass, there must be force. Force is always in the direction of acceleration, and therefore the force is directed toward the center of the circle.

In the above paragraph, we assumed the existence of mass, since all along the discussion has concerned an object spinning around a circle. By definition, an object—that is, an item of matter, rather than an imaginary point—possesses mass. Mass is a measure of inertia, which can be explained by the first law of motion: an object in motion tends to remain in motion, at the same speed and in the same direction (that is, at the same velocity) unless or until some outside force acts on it. This tendency to maintain velocity is inertia. Put another way, it is inertia that causes an object standing still to remain motionless, and

A clothoid loop in a roller coaster in Haines City, Florida. At the top of a loop, you feel lighter than normal; at the bottom, you feel heavier. *(The Purcell Team/Corbis. Reproduced by permission.)*

likewise, it is inertia which dictates that a moving object will "try" to keep moving.

Centripetal Force

Now that we have established the existence of a force in rotational motion, it is possible to give it a name: centripetal force, or the force that causes an object in uniform circular motion to move toward the center of the circular path. This is not a "new" kind of force; it is merely force as applied in circular or rotational motion, and it is absolutely essential. Hence, physicists speak of a "centripetal force requirement": in the absence of centripetal force, an object simply cannot turn. Instead, it will move in a straight line.

The Latin roots of centripetal together mean "seeking the center." What, then, of centrifugal, a word that means "fleeing the center"? It would be correct to say that there is such a thing as centrifugal motion; but centrifugal force is quite a different matter. The difference between centripetal force and a mere centrifugal tendency—a result of inertia rather than of force—can be explained by referring to a familiar example.

Real-Life Applications

Riding in a Car

When you are riding in a car and the car accelerates, your body tends to move backward against the seat. Likewise, if the car stops suddenly, your body tends to move forward, in the direction of the dashboard. Note the language here: "tends to move" rather than "is pushed." To say that something is pushed would imply that a force has been applied, yet what is at work here is not a force, but inertia—the tendency of an object in motion to remain in motion, and an object at rest to remain at rest.

A car that is not moving is, by definition, at rest, and so is the rider. Once the car begins moving, thus experiencing a change in velocity, the rider's body still tends to remain in the fixed position. Hence, it is not a force that has pushed the rider backward against the seat; rather, force has pushed the car forward, and the seat moves up to meet the rider's back. When stopping, once again, there is a sudden change in velocity from a certain value down to zero. The rider, meanwhile, is continuing to move forward due to inertia, and thus, his or her body has a tendency to keep moving in the direction of the now-stationary dashboard.

This may seem a bit too simple to anyone who has studied inertia, but because the human mind has such a strong inclination to perceive inertia as a force in itself, it needs to be clarified in the most basic terms. This habit is similar to the experience you have when sitting in a vehicle that is standing still, while another vehicle alongside moves backward. In the first split-second of awareness, your mind tends to interpret the backward motion of the other car as forward motion on the part of the car in which you are sitting—even though your own car is standing still.

Now we will consider the effects of centripetal force, as well as the illusion of centrifugal force. When a car turns to the left, it is undergoing a form of rotation, describing a 90°-angle or one-quarter of a circle. Once again, your body experiences inertia, since it was in motion along with the car at the beginning of the turn, and thus you tend to move forward. The car, at the same time, has largely overcome its own inertia and moved into the leftward turn. Thus the car door itself is moving to the left. As the door meets the right side of your body, you have the sensation of being pushed outward against the door, but in fact what has happened is that the door has moved inward.

The illusion of centrifugal force is so deeply ingrained in the popular imagination that it warrants further discussion below. But while on the subject of riding in an automobile, we need to examine another illustration of centripetal force. It should be noted in this context that for a car to make a turn at all, there must be friction between the tires and the road. Friction is the force that resists motion when the surface of one object comes into contact with the surface of another; yet ironically, while opposing motion, friction also makes relative motion possible.

Suppose, then, that a driver applies the brakes while making a turn. This now adds a force tangential, or at a right angle, to the centripetal force. If this force is greater than the centripetal force—that is, if the car is moving too fast—the vehicle will slide forward rather than making the turn. The results, as anyone who has ever been in this situation will attest, can be disastrous.

The above highlights the significance of the centripetal force requirement: without a sufficient degree of centripetal force, an object simply cannot turn. Curves are usually banked to maximize centripetal force, meaning that the roadway tilts inward in the direction of the curve. This banking causes a change in velocity, and hence, in acceleration, resulting in an additional quantity known as reaction force, which provides the vehicle with the centripetal force necessary for making the turn.

The formula for calculating the angle at which a curve should be banked takes into account the car's speed and the angle of the curve, but does not include the mass of the vehicle itself. As a result, highway departments post signs stating the speed at which vehicles should make the turn, but these signs do not need to include specific statements regarding the weight of given models.

The Centrifuge

To return to the subject of "centrifugal force"—which, as noted earlier, is really just centrifugal motion—you might ask, "If there is no such thing as centrifugal force, how does a centrifuge work?" Used widely in medicine and a variety of sciences, a centrifuge is a device that separates particles within a liquid. One application, for instance, is to separate red blood cells from plasma.

Typically a centrifuge consists of a base; a rotating tube perpendicular to the base; and two vials attached by movable centrifuge arms to the rotating tube. The movable arms are hinged at the top of the rotating tube, and thus can move upward at an angle approaching 90° to the tube. When the tube begins to spin, centripetal force pulls the material in the vials toward the center.

Materials that are denser have greater inertia, and thus are less responsive to centripetal force. Hence, they seem to be pushed outward, but in fact what has happened is that the less dense material has been pulled inward. This leads to the separation of components, for instance, with plasma on the top and red blood cells on the bottom. Again, the plasma is not as dense, and thus is more easily pulled toward the center of rotation, whereas the red blood cells respond less, and consequently remain on the bottom.

The centrifuge was invented in 1883 by Carl de Laval (1845-1913), a Swedish engineer, who used it to separate cream from milk. During the 1920s, the chemist Theodor Svedberg (1884-1971), who was also Swedish, improved on Laval's work to create the ultracentrifuge, used for separating very small particles of similar weight.

In a typical ultracentrifuge, the vials are no larger than 0.2 in (0.6 cm) in diameter, and these may rotate at speeds of up to 230,000 revolutions per minute. Most centrifuges in use by industry rotate in a range between 1,000 and 15,000 revolutions per minute, but others with scientific applications rotate at a much higher rate, and can produce a force more than 25,000 times as great as that of gravity.

In 1994, researchers at the University of Colorado created a sort of super-centrifuge for simulating stresses applied to dams and other large structures. The instrument has just one centrifuge arm, measuring 19.69 ft (6 m), attached to which is a swinging basket containing a scale model of the structure to be tested. If the model is 1/50 the size of the actual structure, then the centrifuge is set to create a centripetal force 50 times that of gravity.

The Colorado centrifuge has also been used to test the effects of explosions on buildings. Because the combination of forces—centripetal, gravity, and that of the explosion itself—is so great, it takes a very small quantity of explosive to measure the effects of a blast on a model of the building.

Industrial uses of the centrifuge include that for which Laval invented it—separation of cream from milk—as well as the separation of impurities from other substances. Water can be removed from oil or jet fuel with a centrifuge, and likewise, waste-management agencies use it to separate solid materials from waste water prior to purifying the water itself.

Closer to home, a washing machine on spin cycle is a type of centrifuge. As the wet clothes spin, the water in them tends to move outward, separating from the clothes themselves. An even simpler, more down-to-earth centrifuge can be created by tying a fairly heavy weight to a rope and swinging it above one's head: once again, the weight behaves as though it were pushed outward, though in fact, it is only responding to inertia.

Roller Coasters and Centripetal Force

People ride roller coasters, of course, for the thrill they experience, but that thrill has more to do with centripetal force than with speed. By the late twentieth century, roller coasters capable of speeds above 90 MPH (144 km/h) began to appear in amusement parks around America; but prior to that time, the actual speeds of a roller coaster were not particularly impressive. Seldom, if ever, did they exceed that of a car moving down the highway. On the other hand, the acceleration and centripetal force generated on a roller coaster are high, conveying a sense of weightlessness (and sometimes the opposite of weightlessness) that is memorable indeed.

Few parts of a roller coaster ride are straight and flat—usually just those segments that mark the end of one ride and the beginning of another. The rest of the track is generally composed of dips and hills, banked turns, and in some cases, clothoid loops. The latter refers to a geometric shape known as a clothoid, rather like a teardrop upside-down.

Because of its shape, the clothoid has a much smaller radius at the top than at the bottom—a key factor in the operation of the roller coaster ride through these loops. In days past, roller-coaster designers used perfectly circular loops, which allowed cars to enter them at speeds that were too high, built too much force and resulted in injuries for riders. Eventually, engineers recognized the clothoid as a means of providing a safe, fun ride.

KEY TERMS

ACCELERATION: A change in velocity.

CENTRIFUGAL: A term describing the tendency of objects in uniform circular motion to move away from the center of the circular path. Though the term "centrifugal force" is often used, it is inertia, rather than force, that causes the object to move outward.

CENTRIPETAL FORCE: The force that causes an object in uniform circular motion to move toward the center of the circular path.

INERTIA: The tendency of an object in motion to remain in motion, and of an object at rest to remain at rest.

MASS: A measure of inertia, indicating the resistance of an object to a change in its motion—including a change in velocity.

SCALAR: A quantity that possesses only magnitude, with no specific direction. Mass, time, and speed are all scalars. A scalar is contrasted with a vector.

SPEED: The rate at which the position of an object changes over a given period of time.

TANGENTIAL: Movement along a tangent, or a line that touches a circle at just one point and does not intersect the circle.

UNIFORM CIRCULAR MOTION: The motion of an object around the center of a circle in such a manner that speed is constant or unchanging.

VECTOR: A quantity that possesses both magnitude and direction. Velocity, acceleration, and weight (which involves the downward acceleration due to gravity) are examples of vectors. It is contrasted with a scalar.

VELOCITY: The speed of an object in a particular direction.

As you move into the clothoid loop, then up, then over, and down, you are constantly changing position. Speed, too, is changing. On the way up the loop, the roller coaster slows due to a decrease in kinetic energy, or the energy that an object possesses by virtue of its movement. At the top of the loop, the roller coaster has gained a great deal of potential energy, or the energy an object possesses by virtue of its position, and its kinetic energy is at zero. But once it starts going down the other side, kinetic energy—and with it speed—increases rapidly once again.

Throughout the ride, you experience two forces, gravity, or weight, and the force (due to motion) of the roller coaster itself, known as normal force. Like kinetic and potential energy—which rise and fall correspondingly with dips and hills—normal force and gravitational force are locked in a sort of "competition" throughout the roller-coaster rider. For the coaster to have its proper effect, normal force must exceed that of gravity in most places.

The increase in normal force on a roller-coaster ride can be attributed to acceleration and centripetal motion, which cause you to experience something other than gravity. Hence, at the top of a loop, you feel lighter than normal, and at the bottom, heavier. In fact, there has been no real change in your weight: it is, like the idea of "centrifugal force" discussed earlier, a matter of perception.

WHERE TO LEARN MORE

Aylesworth, Thomas G. *Science at the Ball Game.* New York: Walker, 1977.

Beiser, Arthur. *Physics,* 5th ed. Reading, MA: Addison-Wesley, 1991.

Buller, Laura and Ron Taylor. *Forces of Nature.* Illustrations by John Hutchinson and Stan North. New York: Marshall Cavendish, 1990.

"Centrifugal Force—Rotational Motion." National Aeronautics and Space Administration (Web site). <http://observe.ivv.nasa.gov/nasa/space/centrifugal/centrifugal3.html> (March 5, 2001).

"Circular and Satellite Motion" (Web site). <http://www.glenbrook.k12.il.us/gbssci/phys/Class/circles/circtoc.html> (March 5, 2001).

Cobb, Vicki. *Why Doesn't the Earth Fall Up? And Other Not Such Dumb Questions About Motion.* Illustrated by Ted Enik. New York: Lodestar Books, 1988.

Lefkowitz, R. J. *Push! Pull! Stop! Go! A Book About Forces and Motion.* Illustrated by June Goldsborough. New York: Parents' Magazine Press, 1975.

"Rotational Motion." Physics Department, University of Guelph (Web site). <http://www.physics.uoguelph.ca/tutorials/torque/> (March 4, 2001).

Schaefer, Lola M. *Circular Movement.* Mankato, MN: Pebble Books, 2000.

Snedden, Robert. *Forces.* Des Plaines, IL: Heinemann Library, 1999.

Whyman, Kathryn. *Forces in Action.* New York: Gloucester Press, 1986.

FRICTION

CONCEPT

Friction is the force that resists motion when the surface of one object comes into contact with the surface of another. In a machine, friction reduces the mechanical advantage, or the ratio of output to input: an automobile, for instance, uses one-quarter of its energy on reducing friction. Yet, it is also friction in the tires that allows the car to stay on the road, and friction in the clutch that makes it possible to drive at all. From matches to machines to molecular structures, friction is one of the most significant phenomena in the physical world.

HOW IT WORKS

The definition of friction as "the force that resists motion when the surface of one object comes into contact with the surface of another" does not exactly identify what it is. Rather, the statement describes the manifestation of friction in terms of how other objects respond. A less sophisticated version of such a definition would explain electricity, for instance, as "the force that runs electrical appliances." The reason why friction cannot be more firmly identified is simple: physicists do not fully understand what it is.

Much the same could be said of force, defined by Sir Isaac Newton's (1642-1727) second law of motion as the product of mass multiplied acceleration. The fact is that force is so fundamental that it defies full explanation, except in terms of the elements that compose it, and compared to force, friction is relatively easy to identify. In fact, friction plays a part in the total force that must be opposed in order for movement to take place in many situations. So, too, does gravity—and gravity, unlike force itself, is much easier to explain. Since gravity plays a role in friction, it is worthwhile to review its essentials.

Newton's first law of motion identifies inertia, a tendency of objects in the physical universe that is sometimes mistaken for friction. When an object is in motion or at rest, the first law states, it will remain in that state at a constant velocity (which is zero for an object at rest) unless or until an outside force acts on it. This tendency to remain in a given state of motion is inertia.

Inertia is not a force: on the contrary, a very small quantity of force may accelerate an object, thus overcoming its inertia. Inertia is, however, a component of force, since mass is a measure of inertia. In the case of gravitational force, mass is multiplied by the acceleration due to gravity, which is equal to 32 ft (9.8 m)/sec^2. People in everyday life are familiar with another term for gravitational force: weight.

Weight, in turn, is an all-important factor in friction, as revealed in the three laws governing the friction between an object at rest and the surface on which it sits. According to the first of these, friction is proportional to the weight of the object. The second law states that friction is not determined by the surface area of the object—that is, the area that touches the surface on which the object rests. In fact, the contact area between object and surface is a dependant variable, a function of weight.

The second law might seem obvious if one were thinking of a relatively elastic object—say, a garbage bag filled with newspapers sitting on the finished concrete floor of a garage. Clearly as more newspapers are added, thus increasing the weight, its surface area would increase as well. But what if one were to compare a large cardboard box (the kind, for instance, in which televisions or computers are shipped) with an ordinary concrete block of the type used in foundations for residential construction? Obviously, the block has more friction against the concrete floor; but at the same time, it is clear that despite its greater weight, the block has less surface area than the box. How can this be?

The answer is that "surface area" is quite literally more than meets the eye. Friction itself occurs at a level invisible to the naked eye, and involves the adhesive forces between molecules on surfaces pushed together by the force of weight. This is similar to the manner in which, when viewed through a high-powered lens, two complementary patches of Velcro™ are revealed as a forest of hooks on the one hand, and a sea of loops on the other.

On a much more intensified level, that of molecular structure, the surfaces of objects appear as mountains and valleys. Nothing, in fact, is smooth when viewed on this scale, and hence, from a molecular perspective, it becomes clear that two objects in contact actually touch one another only in places. An increase of weight, however, begins pushing objects together, causing an increase in the actual—that is, the molecular—area of contact. Hence area of contact is proportional to weight.

Just as the second law regarding friction states that surface area does not determine friction (but rather, weight determines surface area), the third law holds that friction is independent of the speed at which an object is moving along a surface—provided that speed is not zero. The reason for this provision is that an object with no speed (that is, one standing perfectly still) is subject to the most powerful form of friction, static friction.

The latter is the friction that an object at rest must overcome to be set in motion; however, this should not be confused with inertia, which is relatively easy to overcome through the use of force. Inertia, in fact, is far less complicated than static friction, involving only mass rather than weight. Nor is inertia affected by the composition of the materials touching one another.

As stated earlier, friction is proportional to weight, which suggests that another factor is involved. And indeed there is another factor, known as coefficient of friction. The latter, designated by the Greek letter mu (μ), is constant for any two types of surface in contact with one another, and provides a means of comparing the friction between them to that between other surfaces. For instance, the coefficient of static friction for wood on wood is 0.5; but for metal on metal with lubrication in between, it is only 0.03. A rubber tire on dry concrete yields the highest coefficient of static friction, 1.0, which is desirable in that particular situation.

The raised tread on an automobile's tires, coupled with the roughened road surface, provides sufficient friction for the driver to be able to turn the car and to stop. *(Photograph by Martyn Goddard/Corbis. Reproduced by permission.)*

Coefficients are much lower for the second type of friction, sliding friction, the frictional resistance experienced by a body in motion. Whereas the earlier figures measured the relative resistance to putting certain objects into motion, the sliding-friction coefficient indicates the relative resistance against those objects once they are moving. To use the same materials mentioned above, the coefficient of sliding friction for wood on wood is 0.3; for two lubricated metals 0.03 (no change); and for a rubber tire on dry concrete 0.7.

Finally, there is a third variety of friction, one in which coefficients are so low as to be neg-

ACTOR JACK BUCHANAN LIT THIS MATCH USING A SIMPLE DISPLAY OF FRICTION: DRAGGING THE MATCH AGAINST THE BACK OF THE MATCHBOX. *(Photograph by Hulton-Deutsch Collection/Corbis. Reproduced by permission.)*

ligible: rolling friction, or the frictional resistance that a wheeled object experiences when it rolls over a relatively smooth, flat surface. In ideal circumstances, in fact, there would be absolutely no resistance between a wheel and a road. However, there exists no ideal—that is, perfectly rigid—wheel or road; both objects "give" in response to the other, the wheel by flattening somewhat and the road by experiencing indentation.

Up to this point, coefficients of friction have been discussed purely in comparative terms, but in fact, they serve a function in computing frictional force—that is, the force that must be overcome to set an object in motion, or to keep it in motion. Frictional force is equal to the coefficient of friction multiplied by normal force—that is, the perpendicular force that one object or surface exerts on another. On a horizontal plane, normal force is equal to gravity and hence weight. In this equation, the coefficient of friction establishes a limit to frictional force: in order to move an object in a given situation, one must exert a force in excess of the frictional force that keeps it from moving.

REAL-LIFE APPLICATIONS

SELF-MOTIVATION THROUGH FRICTION

Friction, in fact, always opposes movement; why, then, is friction necessary—as indeed it is—for walking, and for keeping a car on the road? The answer relates to the differences between friction and inertia alluded to earlier. In situations of static friction, it is easy to see how a person might confuse friction with inertia, since both serve to keep an object from moving. In situations of sliding or rolling friction, however, it is easier to see the difference between friction and inertia.

Whereas friction is always opposed to movement, inertia is not. When an object is not moving, its inertia does oppose movement—but when the object is in motion, then inertia resists stopping. In the absence of friction or other forces, inertia allows an object to remain in motion forever. Imagine a hockey player hitting a puck across a very, very large rink. Because ice has a much smaller coefficient of friction with regard to the puck than does dirt or asphalt, the puck will travel much further. Still, however, the ice has some friction, and, therefore, the puck will come to a stop at some point.

Now suppose that instead of ice, the surface and objects in contact with it were friction-free, possessing a coefficient of zero. Then what would happen if the player hit the puck? Assuming for the purposes of this thought experiment, that the rink covered the entire surface of Earth, it would travel and travel and travel, ultimately going around the planet. It would never stop, because there would be no friction to stop it, and therefore inertia would have free rein.

The same would be true if one were to firmly push the hockey player with enough force (small in the absence of friction) to set him in motion: he would continue riding around the planet indefinitely, borne by his skates. But what if instead of being set in motion, the hockey player tried to set himself in motion by the action of his skates against the rink's surface?

He would be unable to move even a hair's breadth. The fact is that while static friction opposes the movement of an object from a position of rest to a state of motion, it may—assuming it can be overcome to begin motion at all—be indispensable to that movement. As with the skater in perpetual motion across the rink, the absence of friction means that inertia is "in control;" with friction, however, it is possible to overcome inertia.

Friction in Driving a Car

The same principle applies to a car's tires: if they were perfectly smooth—and, to make matters worse, the road were perfectly smooth as well—the vehicle would keep moving forward when the driver attempted to stop. For this reason, tires are designed with raised tread to maintain a high degree of friction, gripping the road tightly and dispersing water when the roadway is wet.

The force of friction, in fact, pervades the entire operation of a car, and makes it possible for the tires themselves to turn. The turning force, or torque, that the driver exerts on the steering wheel is converted into forces that drive the tires, and these in turn use friction to provide traction. Between steering wheel and tires, of course, are a number of steps, with the engine rotating the crankshaft and transmitting power to the clutch, which applies friction to translate the motion of the crankshaft to the gearbox.

When the driver of a car with a manual transmission presses down on the clutch pedal, this disengages the clutch itself. A clutch is a circular mechanism containing (among other things) a pressure plate, which lifts off the clutch plate. As a result, the flywheel—the instrument that actually transmits force from the crankshaft—is disengaged from the transmission shaft. With the clutch thus disengaged, the driver changes gears, and after the driver releases the clutch pedal, springs return the pressure plate and the clutch plate to their place against the flywheel. The flywheel then turns the transmission shaft.

Controlled friction in the clutch makes this operation possible; likewise the synchromesh within the gearbox uses friction to bring the gearwheels into alignment. This is a complicated process, but at the heart of it is an engagement of gear teeth in which friction forces them to come to the same speed.

Friction is also essential to stopping a car—not just with regard to the tires, but also with respect to the brakes. Whether they are disk brakes or drum brakes, two elements must come together with a force more powerful than the engine's, and friction provides that needed force. In disk brakes, brake pads apply friction to both sides of the spinning disks, and in drum brakes, brake shoes transmit friction to the inside of a spinning drum. This braking force is then transmitted to the tires, which apply friction to the road and thus stop the car.

Efficiency and Friction

The automobile is just one among many examples of a machine that could not operate without friction. The same is true of simple machines such as screws, as well as nails, pliers, bolts, and forceps. At the heart of this relationship is a paradox, however, because friction inevitably reduces the efficiency of machines: a car, as noted earlier, exerts fully one-quarter of its power simply on overcoming the force of friction both within its engine and from air resistance as it travels down the road.

In scientific terms, efficiency or mechanical advantage is measured by the ratio of force output to force input. Clearly, in most situations it is ideal to maximize output and minimize input, and over the years inventors have dreamed of creating a mechanism—a perpetual motion machine—to do just that. In this idealized machine, one would apply a certain amount of energy to set it into operation, and then it would never stop; hence the ratio of output to input would be nearly infinite.

Unfortunately, the perpetual motion machine is a dream every bit as elusive as the mythical Fountain of Youth. At least this is true on Earth, where friction will always cause a system to lose kinetic energy, or the energy of movement. No matter what the design, the machine will eventually lose energy and stop; however, this is not true in outer space, where friction is very small—though it still exists. In space it might truly be possible to set a machine in motion and let inertia do the rest; thus perhaps perpetual motion actually is more than a dream.

It should also be noted that mechanical advantage is not always desirable. A screw is a highly inefficient machine: one puts much more force into screwing it in than the screw will exert once it is in place. Yet this is exactly the purpose of a screw: an "efficient" one, or one that worked its way back out of the place into which it had been screwed, would in fact be of little use.

Once again, it is friction that provides a screw with its strangely efficient form of inefficiency. Nonetheless, friction, in spite of the advantages discussed above, is as undesirable as it is desirable. With friction, there is always something lost; however, there is a physical law that energy does not simply disappear; it just changes form. In the case of friction, the energy that could go to moving the machine is instead translated into sound—or even worse, heat.

When Sparks Fly

In movement involving friction, molecules vibrate, bringing about a rise in temperature. This can be easily demonstrated by simply rubbing one's hands together quickly, as a person is apt to do when cold: heat increases. For a machine composed of metal parts, this increase in temperature can be disastrous, leading to serious wear and damage. This is why various forms of lubricant are applied to systems subject to friction.

An automobile uses grease and oil, as well as ball bearings, which are tiny uniform balls of metal that imitate the behavior of oil-based substances on a large scale. In a molecule of oil—whether it is a petroleum-related oil or the type of oil that comes from living things—positive and negative electrical charges are distributed throughout the molecule. By contrast, in water the positive charges are at one end of the molecule and the negative at the other. This creates a tight bond as the positive end of one water molecule adheres to the negative end of another. With oil, the relative absence of attraction between molecules means that each is in effect a tiny ball separate from the others. The ball-like molecules "roll" between metal elements, providing the buffer necessary to reduce friction.

Yet for every statement one can make concerning friction, there is always another statement with which to counter it. Earlier it was noted that the wheel, because it reduced friction greatly, provided an enormous technological boost to societies. Yet long before the wheel—hundreds of thousands of years ago—an even more important technological breakthrough occurred when humans made a discovery that depended on maximizing friction: fire, or rather the means of making fire. Unlike the wheel, fire occurs in nature, and can spring from a number of causes, but when human beings harnessed the means of making fire on their own, they had to rely on the heat that comes from friction.

By the early nineteenth century, inventors had developed an easy method of creating fire by using a little stick with a phosphorus tip. This stick, of course, is known as a match. In a strike-anywhere match, the head contains all the chemicals needed to create a spark. To ignite this type of match, one need only create frictional heat by rubbing it against a surface, such as sandpaper, with a high coefficient of friction.

The chemicals necessary for ignition in safety matches, on the other hand, are dispersed between the match head and a treated strip, usually found on the side of the matchbox or matchbook. The chemicals on the tip and those on the striking surface must come into contact for ignition to occur, but once again, there must be friction between the match head and the striking pad. Water reduces friction with its heavy bond, as it does with a car's tires on a rainy day, which explains why matches are useless when wet.

The Outer Limits of Friction

Clearly friction is a complex subject, and the discoveries of modern physics only promise to add to that complexity. In a February 1999 online article for *Physical Review Focus,* Dana Mackenzie reported that "Engineers hope to make microscopic engines and gears as ordinary in our lives as microscopic circuits are today. But before this dream becomes a reality, they will have to deal with laws of friction that are very different from those that apply to ordinary-sized machines."

The earlier statement that friction is proportional to weight, in fact, applies only in the realm of classical physics. The latter term refers to the studies of physicists up to the end of the nineteenth century, when the concerns were chiefly the workings of large objects whose operations could be discerned by the senses. Modern physics, on the other hand, focuses on atomic and molecular structures, and addresses physical behaviors that could not have been imagined prior to the twentieth century.

According to studies conducted by Alan Burns and others at Sandia National Laboratories in Albuquerque, New Mexico, molecular interactions between objects in very close proximity create a type of friction involving repulsion rather than attraction. This completely upsets the model of friction understood for more than a century, and indicates new frontiers of discovery concerning the workings of friction at a molecular level.

KEY TERMS

ACCELERATION: A change in velocity.

COEFFICIENT OF FRICTION: A figure, constant for a particular pair of surfaces in contact, that can be multiplied by the normal force between them to calculate the frictional force they experience.

FORCE: The product of mass multiplied by acceleration.

FRICTION: The force that resists motion when the surface of one object comes into contact with the surface of another. Varieties including sliding friction, static friction, and rolling friction. The degree of friction between two specific surfaces is proportional to coefficient of friction.

FRICTIONAL FORCE: The force necessary to set an object in motion, or to keep it in motion; equal to normal force multiplied by coefficient of friction.

INERTIA: The tendency of an object in motion to remain in motion, and of an object at rest to remain at rest.

MASS: A measure of inertia, indicating the resistance of an object to a change in its motion—including a change in velocity.

MECHANICAL ADVANTAGE: The ratio of force output to force input in a machine.

NORMAL FORCE: The perpendicular force with which two objects press against one another. On a plane without any incline (which would add acceleration in addition to that of gravity) normal force is the same as weight.

ROLLING FRICTION: The frictional resistance that a circular object experiences when it rolls over a relatively smooth, flat surface. With a coefficient of friction much smaller than that of sliding friction, rolling friction involves by far the least amount of resistance among the three varieties of friction.

SLIDING FRICTION: The frictional resistance experienced by a body in motion. Here the coefficient of friction is greater than that for rolling friction, but less than for static friction.

SPEED: The rate at which the position of an object changes over a given period of time.

STATIC FRICTION: The frictional resistance that a stationary object must overcome before it can go into motion. Its coefficient of friction is greater than that of sliding friction, and thus largest among the three varieties of friction.

VELOCITY: The speed of an object in a particular direction.

WEIGHT: A measure of the gravitational force on an object; the product of mass multiplied by the acceleration due to gravity.

WHERE TO LEARN MORE

Beiser, Arthur. *Physics,* 5th ed. Reading, MA: Addison-Wesley, 1991.

Buller, Laura and Ron Taylor. *Forces of Nature.* Illustrations by John Hutchinson and Stan North. New York: Marshall Cavendish, 1990.

Dixon, Malcolm and Karen Smith. *Forces and Movement.* Mankato, MN: Smart Apple Media, 1998.

"*Friction.*" *How Stuff Works* (Web site). <http://www.howstuffworks.com/search/index.htm?words=friction> (March 8, 2001).

"*Friction and Interactions*" (Web site). <http://www.cord.edu/dept/physics/p128/lecture99_12.html> (March 8, 2001).

Levy, Matthys and Richard Panchyk. *Engineering the City: How Infrastructure Works.* Chicago: Chicago Review Press, 2000.

Macaulay, David. *The New Way Things Work.* Boston: Houghton Mifflin, 1998.

Mackenzie, Dana. *"Friction of Molecules." Physical Review Focus* (Web site). <http://focus.aps.org/v3/st9.html (March 8, 2001).

Rutherford, F. James; Gerald Holton; and Fletcher G. Watson. *Project Physics.* New York: Holt, Rinehart, and Winston, 1981.

Skateboard Science (Web site). <http://www.exploratorium.edu/skateboarding/ (March 8, 2001).

Suplee, Curt. *Everyday Science Explained.* Washington, D.C.: National Geographic Society, 1996.

LAWS OF MOTION

CONCEPT

In all the universe, there are few ideas more fundamental than those expressed in the three laws of motion. Together these explain why it is relatively difficult to start moving, and then to stop moving; how much force is needed to start or stop in a given situation; and how one force relates to another. In their beauty and simplicity, these precepts are as compelling as a poem, and like the best of poetry, they identify something that resonates through all of life. The applications of these three laws are literally endless: from the planets moving through the cosmos to the first seconds of a car crash to the action that takes place when a person walks. Indeed, the laws of motion are such a part of daily life that terms such as inertia, force, and reaction extend into the realm of metaphor, describing emotional processes as much as physical ones.

HOW IT WORKS

The three laws of motion are fundamental to mechanics, or the study of bodies in motion. These laws may be stated in a number of ways, assuming they contain all the components identified by Sir Isaac Newton (1642-1727). It is on his formulation that the following are based:

The Three Laws of Motion

- First law of motion: An object at rest will remain at rest, and an object in motion will remain in motion, at a constant velocity unless or until outside forces act upon it.
- Second law of motion: The net force acting upon an object is a product of its mass multiplied by its acceleration.
- Third law of motion: When one object exerts a force on another, the second object exerts on the first a force equal in magnitude but opposite in direction.

Laws of Man vs. Laws of Nature

These, of course, are not "laws" in the sense that people normally understand that term. Human laws, such as injunctions against stealing or parking in a fire lane, are prescriptive: they state how the world should be. Behind the prescriptive statements of civic law, backing them up and giving them impact, is a mechanism—police, courts, and penalties—for ensuring that citizens obey.

A scientific law operates in exactly the opposite fashion. Here the mechanism for ensuring that nature "obeys" the law comes first, and the "law" itself is merely a descriptive statement concerning evident behavior. With human or civic law, it is clearly possible to disobey: hence, the justice system exists to discourage disobedience. In the case of scientific law, disobedience is clearly impossible—and if it were not, the law would have to be amended.

This is not to say, however, that scientific laws extend beyond their own narrowly defined limits. On Earth, the intrusion of outside forces—most notably friction—prevents objects from behaving perfectly according to the first law of motion. The common-sense definition of friction calls to mind, for instance, the action that a match makes as it is being struck; in its broader scientific meaning, however, friction can be defined as any force that resists relative motion between two bodies in contact.

The cargo bay of the space shuttle Discovery, shown just after releasing a satellite. Once released into the frictionless vacuum around Earth, the satellite will move indefinitely around Earth without need for the motive power of an engine. The planet's gravity keeps it at a fixed height, and at that height, it could theoretically circle Earth forever. *(Corbis. Reproduced by permission.)*

The operations of physical forces on Earth are continually subject to friction, and this includes not only dry bodies, but liquids, for instance, which are subject to viscosity, or internal friction. Air itself is subject to viscosity, which prevents objects from behaving perfectly in accordance with the first law of motion. Other forces, most notably that of gravity, also come into play to stop objects from moving endlessly once they have been set in motion.

The vacuum of outer space presents scientists with the most perfect natural laboratory for testing the first law of motion: in theory, if they were to send a spacecraft beyond Earth's orbital radius, it would continue travelling indefinitely. But even this craft would likely run into another object, such as a planet, and would then be drawn into its orbit. In such a case, however, it can be said that outside forces have acted upon it, and thus the first law of motion stands.

The orbit of a satellite around Earth illustrates both the truth of the first law, as well as the forces that limit it. To break the force of gravity, a powered spacecraft has to propel the satellite into the exosphere. Yet once it has reached the frictionless vacuum, the satellite will move indefinitely around Earth without need for the motive power of an engine—it will get a "free ride," thanks to the first law of motion. Unlike the hypothetical spacecraft described above, however, it will not go spinning into space, because it is still too close to Earth. The planet's gravity keeps it at a fixed height, and at that height, it could theoretically circle Earth forever.

The first law of motion deserves such particular notice, not simply because it is the first law. Nonetheless, it is first for a reason, because it establishes a framework for describing the behavior of an object in motion. The second law identifies a means of determining the force necessary to move an object, or to stop it from moving, and the third law provides a picture of what happens when two objects exert force on one another.

The first law warrants special attention because of misunderstandings concerning it, which spawned a debate that lasted nearly twenty centuries. Aristotle (384-322 B.C.) was the first scientist to address seriously what is now known as the first law of motion, though in fact, that term would not be coined until about two thousand years after his death. As its title suggests, his *Physics* was a seminal work, a book in which Aristotle attempted to give form to, and thus define the territory of, studies regarding the operation of physical processes. Despite the great philosopher's many achievements, however, *Physics* is a highly flawed work, particularly with regard to what became known as his theory of impetus—that is, the phenomena addressed in the first law of motion.

Aristotle's Mistake

According to Aristotle, a moving object requires a continual application of force to keep it moving: once that force is no longer applied, it ceases to move. You might object that, when a ball is in flight, the force necessary to move it has already been applied: a person has thrown the ball, and it is now on a path that will eventually be stopped by the force of gravity. Aristotle, however, would have maintained that the air itself acts as a force to keep the ball in flight, and that when the ball drops—of course he had no concept of "gravity" as such—it is because the force of the air on the ball is no longer in effect.

These notions might seem patently absurd to the modern mind, but they went virtually unchallenged for a thousand years. Then in the sixth century A.D., the Byzantine philosopher Johannes Philoponus (c. 490-570) wrote a critique of *Physics*. In what sounds very much like a precursor to the first law of motion, Philoponus held that a body will keep moving in the absence of friction or opposition.

He further maintained that velocity is proportional to the positive difference between force and resistance—in other words, that the force propelling an object must be greater than the resistance. As long as force exceeds resistance, Philoponus held, a body will remain in motion. This in fact is true: if you want to push a refrigerator across a carpeted floor, you have to exert enough force not only to push the refrigerator, but also to overcome the friction from the floor itself.

The Arab philosophers Ibn Sina (Avicenna; 980-1037) and Ibn Bâjja (Avempace; fl. c. 1100) defended Philoponus's position, and the French scholar Peter John Olivi (1248-1298) became the first Western thinker to critique Aristotle's statements on impetus. Real progress on the subject, however, did not resume until the time of Jean Buridan (1300-1358), a French physicist who went much further than Philoponus had eight centuries earlier.

In his writing, Buridan offered an amazingly accurate analysis of impetus that prefigured all three laws of motion. It was Buridan's position that one object imparts to another a certain amount of power, in proportion to its velocity and mass, that causes the second object to move a certain distance. This, as will be shown below, was amazingly close to actual fact. He was also correct in stating that the weight of an object may increase or decrease its speed, depending on other circumstances, and that air resistance slows an object in motion.

The true breakthrough in understanding the laws of motion, however, came as the result of work done by three extraordinary men whose lives stretched across nearly 250 years. First came Nicolaus Copernicus (1473-1543), who advanced what was then a heretical notion: that Earth, rather than being the center of the universe, revolved around the Sun along with the other planets. Copernicus made his case purely in terms of astronomy, however, with no direct reference to physics.

Galileo's Challenge: The Copernican Model

Galileo Galilei (1564-1642) likewise embraced a heliocentric (Sun-centered) model of the universe—a position the Church forced him to renounce publicly on pain of death. As a result of his censure, Galileo realized that in order to prove the Copernican model, it would be necessary to show why the planets remain in motion as they do. In explaining this, he coined the term inertia to describe the tendency of an object in motion to remain in motion, and an object at rest to remain at rest. Galileo's observations, in fact, formed the foundation for the laws of motion.

In the years that followed Galileo's death, some of the world's greatest scientific minds became involved in the effort to understand the forces that kept the planets in motion around the Sun. Among them were Johannes Kepler (1571-1630), Robert Hooke (1635-1703), and Edmund Halley (1656-1742). As a result of a dispute between Hooke and Sir Christopher Wren (1632-1723) over the subject, Halley brought the question to his esteemed friend Isaac Newton. As it turned out, Newton had long been considering the possibility that certain laws of motion existed, and these he presented in definitive form in his *Principia* (1687).

The impact of the Newton's book, which included his observations on gravity, was nothing short of breathtaking. For the next three centuries, human imagination would be ruled by the Newtonian framework, and only in the twentieth century would the onset of new ideas reveal its limitations. Yet even today, outside the realm of quantum mechanics and relativity theory—in other words, in the world of everyday experience—Newton's laws of motion remain firmly in place.

REAL-LIFE APPLICATIONS

The First Law of Motion in a Car Crash

It is now appropriate to return to the first law of motion, as formulated by Newton: an object at rest will remain at rest, and an object in motion will remain in motion, at a constant velocity unless or until outside forces act upon it. Examples of this first law in action are literally unlimited.

One of the best illustrations, in fact, involves something completely outside the experience of Newton himself: an automobile. As a car moves down the highway, it has a tendency to remain in motion unless some outside force changes its velocity. The latter term, though it is commonly understood to be the same as speed, is in fact more specific: velocity can be defined as the speed of an object in a particular direction.

In a car moving forward at a fixed rate of 60 MPH (96 km/h), everything in the car—driver, passengers, objects on the seats or in the trunk—is also moving forward at the same rate. If that car then runs into a brick wall, its motion will be stopped, and quite abruptly. But though its motion has stopped, in the split seconds after the crash it is still responding to inertia: rather than bouncing off the brick wall, it will continue plowing into it.

What, then, of the people and objects in the car? They too will continue to move forward in response to inertia. Though the car has been stopped by an outside force, those inside experience that force indirectly, and in the fragment of time after the car itself has stopped, they continue to move forward—unfortunately, straight into the dashboard or windshield.

It should also be clear from this example exactly why seatbelts, headrests, and airbags in automobiles are vitally important. Attorneys may file lawsuits regarding a client's injuries from airbags, and homespun opponents of the seatbelt may furnish a wealth of anecdotal evidence concerning people who allegedly died in an accident because they were wearing seatbelts; nonetheless, the first law of motion is on the side of these protective devices.

The admittedly gruesome illustration of a car hitting a brick wall assumes that the driver has not applied the brakes—an example of an outside force changing velocity—or has done so too late. In any case, the brakes themselves, if applied too abruptly, can present a hazard, and again, the significant factor here is inertia. Like the brick wall, brakes stop the car, but there is nothing to stop the driver and/or passengers. Nothing, that is, except protective devices: the seat belt to keep the person's body in place, the airbag to cushion its blow, and the headrest to prevent whiplash in rear-end collisions.

Inertia also explains what happens to a car when the driver makes a sharp, sudden turn. Suppose you are is riding in the passenger seat of a car moving straight ahead, when suddenly the driver makes a quick left turn. Though the car's tires turn instantly, everything in the vehicle—its frame, its tires, and its contents—is still respond-

When a vehicle hits a wall, as shown here in a crash test, its motion will be stopped, and quite abruptly. But though its motion has stopped, in the split seconds after the crash it is still responding to inertia: rather than bouncing off the brick wall, it will continue plowing into it. *(Photograph by Tim Wright/Corbis. Reproduced by permission.)*

ing to inertia, and therefore "wants" to move forward even as it is turning to the left.

As the car turns, the tires may respond to this shift in direction by squealing: their rubber surfaces were moving forward, and with the sudden turn, the rubber skids across the pavement like a hard eraser on fine paper. The higher the original speed, of course, the greater the likelihood the tires will squeal. At very high speeds, it is possible the car may seem to make the turn "on two wheels"—that is, its two outer tires. It is even possible that the original speed was so high, and the turn so sharp, that the driver loses control of the car.

Here inertia is to blame: the car simply cannot make the change in velocity (which, again, refers both to speed and direction) in time. Even in less severe situations, you are likely to feel that you have been thrown outward against the rider's side door. But as in the car-and-brick-wall illustration used earlier, it is the car itself that first experiences the change in velocity, and thus it responds first. You, the passenger, then, are moving forward even as the car has turned; therefore, rather than being thrown outward, you are simply meeting the leftward-moving door even as you push forward.

From Parlor Tricks to Space Ships

It would be wrong to conclude from the car-related illustrations above that inertia is always harmful. In fact it can help every bit as much as it can potentially harm, a fact shown by two quite different scenarios.

The beneficial quality to the first scenario may be dubious: it is, after all, a mere parlor trick, albeit an entertaining one. In this famous stunt, with which most people are familiar even if they have never seen it, a full table setting is placed on a table with a tablecloth, and a skillful practitioner manages to whisk the cloth out from under the dishes without upsetting so much as a glass. To some this trick seems like true magic, or at least sleight of hand; but under the right conditions, it can be done. (This information, however, carries with it the warning, "Do not try this at home!")

To make the trick work, several things must align. Most importantly, the person doing it has to be skilled and practiced at performing the feat. On a physical level, it is best to minimize the friction between the cloth and settings on the one hand, and the cloth and table on the other. It is also important to maximize the mass (a property that will be discussed below) of the table settings, thus making them resistant to movement. Hence, inertia—which is measured by mass—plays a key role in making the tablecloth trick work.

You might question the value of the tablecloth stunt, but it is not hard to recognize the importance of the inertial navigation system (INS) that guides planes across the sky. Prior to the 1970s, when INS made its appearance, navigation techniques for boats and planes relied on reference to external points: the Sun, the stars, the magnetic North Pole, or even nearby areas of land. This created all sorts of possibilities for error: for instance, navigation by magnet (that is, a compass) became virtually useless in the polar regions of the Arctic and Antarctic.

By contrast, the INS uses no outside points of reference: it navigates purely by sensing the inertial force that results from changes in velocity. Not only does it function as well near the poles as it does at the equator, it is difficult to tamper with an INS, which uses accelerometers in a sealed, shielded container. By contrast, radio signals or radar can be "confused" by signals from the ground—as, for instance, from an enemy unit during wartime.

As the plane moves along, its INS measures movement along all three geometrical axes, and provides a continuous stream of data regarding acceleration, velocity, and displacement. Thanks to this system, it is possible for a pilot leaving California for Japan to enter the coordinates of a half-dozen points along the plane's flight path, and let the INS guide the autopilot the rest of the way.

Yet INS has its limitations, as illustrated by the tragedy that occurred aboard Korean Air Lines (KAL) Flight 007 on September 1, 1983. The plane, which contained 269 people and crew members, departed Anchorage, Alaska, on course for Seoul, South Korea. The route they would fly was an established one called "R-20," and it appears that all the information regarding their flight plan had been entered correctly in the plane's INS.

This information included coordinates for internationally recognized points of reference, actually just spots on the northern Pacific with names such as NABIE, NUKKS, NEEVA, and so on, to NOKKA, thirty minutes east of Japan. Yet, just after passing the fishing village of Bethel, Alaska, on the Bering Sea, the plane started to veer off course, and ultimately wandered into Soviet airspace over the Kamchatka Peninsula and later Sakhalin Island. There a Soviet Su-15 shot it down, killing all the plane's passengers.

In the aftermath of the Flight 007 shootdown, the Soviets accused the United States and South Korea of sending a spy plane into their airspace. (Among the passengers was Larry McDonald, a staunchly anti-Communist Congressman from Georgia.) It is more likely, however, that the tragedy of 007 resulted from errors in navigation which probably had something to do with the INS. The fact is that the R-20 flight plan had been designed to keep aircraft well out of Soviet airspace, and at the time KAL 007 passed over Kamchatka, it should have been 200 mi (320 km) to the east—over the Sea of Japan.

Among the problems in navigating a transpacific flight is the curvature of the Earth, combined with the fact that the planet continues to rotate as the aircraft moves. On such long flights, it is impossible to "pretend," as on a short flight, that Earth is flat: coordinates have to be adjusted for the rounded surface of the planet. In addition, the flight plan must take into account that (in the case of a flight from California to Japan), Earth is moving eastward even as the plane moves westward. The INS aboard KAL 007 may simply have failed to correct for these factors, and thus the error compounded as the plane moved further. In any case, INS will eventually be rendered obsolete by another form of navigation technology: the global positioning satellite (GPS) system.

UNDERSTANDING INERTIA

From examples used above, it should be clear that inertia is a more complex topic than you might immediately guess. In fact, inertia as a process is rather straightforward, but confusion regarding its meaning has turned it into a complicated subject.

In everyday terminology, people typically use the word inertia to describe the tendency of a stationary object to remain in place. This is particularly so when the word is used metaphorically: as suggested earlier, the concept of inertia, like numerous other aspects of the laws of motion, is often applied to personal or emotional processes as much as the physical. Hence, you could say, for instance, "He might have changed professions and made more money, but inertia kept him at his old job." Yet you could just as easily say, for example, "He might have taken a vacation, but inertia kept him busy." Because of the misguided way that most people use the term, it is easy to forget that "inertia" equally describes a tendency toward movement or nonmovement: in terms of Newtonian mechanics, it simply does not matter.

The significance of the clause "unless or until outside forces act upon it" in the first law indicates that the object itself must be in equilibrium—that is, the forces acting upon it must be balanced. In order for an object to be in equilibrium, its rate of movement in a given direction must be constant. Since a rate of movement equal to 0 is certainly constant, an object at rest is in equilibrium, and therefore qualifies; but also, any object moving in a constant direction at a constant speed is also in equilibrium.

The Second Law: Force, Mass, Acceleration

As noted earlier, the first law of motion deserves special attention because it is the key to unlocking the other two. Having established in the first law the conditions under which an object in motion will change velocity, the second law provides a measure of the force necessary to cause that change.

Understanding the second law requires defining terms that, on the surface at least, seem like a matter of mere common sense. Even inertia requires additional explanation in light of terms related to the second law, because it would be easy to confuse it with momentum.

The measure of inertia is mass, which reflects the resistance of an object to a change in its motion. Weight, on the other hand, measures the gravitational force on an object. (The concept of force itself will require further definition shortly.) Hence a person's mass is the same everywhere in the universe, but their weight would differ from planet to planet.

This can get somewhat confusing when you attempt to convert between English and metric units, because the pound is a unit of weight or force, whereas the kilogram is a unit of mass. In fact it would be more appropriate to set up kilograms against the English unit called the slug (equal to 14.59 kg), or to compare pounds to the metric unit of force, the newton (N), which is equal to the acceleration of one meter per second per second on an object of 1 kg in mass.

Hence, though many tables of weights and measures show that 1 kg is equal to 2.21 lb, this is only true at sea level on Earth. A person with a mass of 100 kg on Earth would have the same mass on the Moon; but whereas he might weigh 221 lb on Earth, he would be considerably lighter on the Moon. In other words, it would be much easier to lift a 221-lb man on the Moon than on Earth, but it would be no easier to push him aside.

To return to the subject of momentum, whereas inertia is measured by mass, momentum is equal to mass multiplied by velocity. Hence momentum, which Newton called "quantity of motion," is in effect inertia multiplied by velocity. Momentum is a subject unto itself; what matters here is the role that mass (and thus inertia) plays in the second law of motion.

According to the second law, the net force acting upon an object is a product of its mass multiplied by its acceleration. The latter is defined as a change in velocity over a given time interval: hence acceleration is usually presented in terms of "feet (or meters) per second per second"—that is, feet or meters per second squared. The acceleration due to gravity is 32 ft (9.8 m) per second per second, meaning that as every second passes, the speed of a falling object is increasing by 32 ft (9.8 m) per second.

The second law, as stated earlier, serves to develop the first law by defining the force necessary to change the velocity of an object. The law was integral to the confirming of the Copernican model, in which planets revolve around the Sun. Because velocity indicates movement in a single (straight) direction, when an object moves in a curve—as the planets do around the Sun—it is by definition changing velocity, or accelerating. The fact that the planets, which clearly possessed mass, underwent acceleration meant that some force must be acting on them: a gravitational pull exerted by the Sun, most massive object in the solar system.

Gravity is in fact one of four types of force at work in the universe. The others are electromagnetic interactions, and "strong" and "weak" nuclear interactions. The other three were unknown to Newton—yet his definition of force is still applicable. Newton's calculation of gravitational force (which, like momentum, is a subject unto itself) made it possible for Halley to determine that the comet he had observed in

1682—the comet that today bears his name—would reappear in 1758, as indeed it has for every 75–76 years since then. Today scientists use the understanding of gravitational force imparted by Newton to determine the exact altitude necessary for a satellite to remain stationary above the same point on Earth's surface.

The second law is so fundamental to the operation of the universe that you seldom notice its application, and it is easiest to illustrate by examples such as those above—of astronomers and physicists applying it to matters far beyond the scope of daily life. Yet the second law also makes it possible, for instance, to calculate the amount of force needed to move an object, and thus people put it into use every day without knowing that they are doing so.

The Third Law: Action and Reaction

As with the second law, the third law of motion builds on the first two. Having defined the force necessary to overcome inertia, the third law predicts what will happen when one force comes into contact with another force. As the third law states, when one object exerts a force on another, the second object exerts on the first a force equal in magnitude but opposite in direction.

Unlike the second law, this one is much easier to illustrate in daily life. If a book is sitting on a table, that means that the book is exerting a force on the table equal to its mass multiplied by its rate of acceleration. Though it is not moving, the book is subject to the rate of gravitational acceleration, and in fact force and weight (which is defined as mass multiplied by the rate of acceleration due to gravity) are the same. At the same time, the table pushes up on the book with an exactly equal amount of force—just enough to keep it stationary. If the table exerted more force that the book—in other words, if instead of being an ordinary table it were some sort of pneumatic press pushing upward—then the book would fly off the table.

There is no such thing as an unpaired force in the universe. The table rests on the floor just as the book rests on it, and the floor pushes up on the table with a force equal in magnitude to that with which the table presses down on the floor. The same is true for the floor and the supporting beams that hold it up, and for the supporting beams and the foundation of the building, and the building and the ground, and so on.

These pairs of forces exist everywhere. When you walk, you move forward by pushing backward on the ground with a force equal to your mass multiplied by your rate of downward gravitational acceleration. (This force, in other words, is the same as weight.) At the same time, the ground actually pushes back with an equal force. You do not perceive the fact that Earth is pushing you upward, simply because its enormous mass makes this motion negligible—but it does push.

If you were stepping off of a small unmoored boat and onto a dock, however, something quite different would happen. The force of your leap to the dock would exert an equal force against the boat, pushing it further out into the water, and as a result, you would likely end up in the water as well. Again, the reaction is equal and opposite; the problem is that the boat in this illustration is not fixed in place like the ground beneath your feet.

Differences in mass can result in apparently different reactions, though in fact the force is the same. This can be illustrated by imagining a mother and her six-year-old daughter skating on ice, a relatively frictionless surface. Facing one another, they push against each other, and as a result each moves backward. The child, of course, will move backward faster because her mass is less than that of her mother. Because the force they exerted is equal, the daughter's acceleration is greater, and she moves farther.

Ice is not a perfectly frictionless surface, of course: otherwise, skating would be impossible. Likewise friction is absolutely necessary for walking, as you can illustrate by trying to walk on a perfectly slick surface—for instance, a skating rink covered with oil. In this situation, there is still an equally paired set of forces—your body presses down on the surface of the ice with as much force as the ice presses upward—but the lack of friction impedes the physical process of pushing off against the floor.

It will only be possible to overcome inertia by recourse to outside intervention, as for instance if someone who is not on the ice tossed out a rope attached to a pole in the ground. Alternatively, if the person on the ice were carrying a heavy load of rocks, it would be possible to move by throwing the rocks backward. In this situation, you are exerting force on the rock, and this

KEY TERMS

ACCELERATION: A change in velocity over a given time period.

EQUILIBRIUM: A situation in which the forces acting upon an object are in balance.

FRICTION: Any force that resists the motion of body in relation to another with which it is in contact.

INERTIA: The tendency of an object in motion to remain in motion, and of an object at rest to remain at rest.

MASS: A measure of inertia, indicating the resistance of an object to a change in its motion—including a change in velocity. A kilogram is a unit of mass, whereas a pound is a unit of weight. The mass of an object remains the same throughout the universe, whereas its weight is a function of gravity on any given planet.

MECHANICS: The study of bodies in motion.

MOMENTUM: The product of mass multiplied by velocity.

SPEED: The rate at which the position of an object changes over a given period of time.

VELOCITY: The speed of an object in a particular direction.

VISCOSITY: The internal friction in a fluid that makes it resistant to flow.

WEIGHT: A measure of the gravitational force on an object. A pound is a unit of weight, whereas a kilogram is a unit of mass. Weight thus would change from planet to planet, whereas mass remains constant throughout the universe.

backward force results in a force propelling the thrower forward.

This final point about friction and movement is an appropriate place to close the discussion on the laws of motion. Where walking or skating are concerned—and in the absence of a bag of rocks or some other outside force—friction is necessary to the action of creating a backward force and therefore moving forward. On the other hand, the absence of friction would make it possible for an object in movement to continue moving indefinitely, in line with the first law of motion. In either case, friction opposes inertia.

The fact is that friction itself is a force. Thus, if you try to slide a block of wood across a floor, friction will stop it. It is important to remember this, lest you fall into the fallacy that bedeviled Aristotle's thinking and thus confused the world for many centuries. The block did not stop moving because the force that pushed it was no longer being applied; it stopped because an opposing force, friction, was greater than the force that was pushing it.

WHERE TO LEARN MORE

Ardley, Neil. *The Science Book of Motion.* San Diego: Harcourt Brace Jovanovich, 1992.

Beiser, Arthur. *Physics,* 5th ed. Reading, MA: Addison-Wesley, 1991.

Chase, Sara B. *Moving to Win: The Physics of Sports.* New York: Messner, 1977.

Fleisher, Paul. *Secrets of the Universe: Discovering the Universal Laws of Science.* Illustrated by Patricia A. Keeler. New York: Atheneum, 1987.

"The Laws of Motion." How It Flies (Web site). <http://www.monmouth.com/~jsd/how/htm/motion.html> (February 27, 2001).

Newton, Isaac (translated by Andrew Motte, 1729). *The Principia* (Web site). <http://members.tripod.com/~gravitee/principia.html> (February 27, 2001).

Newton's Laws of Motion (Web site). <http://www.glenbrook.k12.il.us/gbssci/phys/Class/newtlaws/newtloc.html> (February 27, 2001).

"*Newton's Laws of Motion." Dryden Flight Research Center, National Aeronautics and Space Administration (NASA)* (Web site). <http://www.dfrc.nasa.gov/trc/saic/newton.html> (February 27, 2001).

"*Newton's Laws of Motion: Movin' On." Beyond Books* (Web site). <http://www.beyondbooks.com/psc91/4.asp> (February 27, 2001).

Roberts, Jeremy. *How Do We Know the Laws of Motion?* New York: Rosen, 2001.

Suplee, Curt. *Everyday Science Explained.* Washington, D.C.: National Geographic Society, 1996.

GRAVITY AND GRAVITATION

CONCEPT

Gravity is, quite simply, the force that holds together the universe. People are accustomed to thinking of it purely in terms of the gravitational pull Earth exerts on smaller bodies—a stone, a human being, even the Moon—or perhaps in terms of the Sun's gravitational pull on Earth. In fact, everything exerts a gravitational attraction toward everything else, an attraction commensurate with the two body's relative mass, and inversely related to the distance between them. The earliest awareness of gravity emerged in response to a simple question: why do objects fall when released from any restraining force? The answers, which began taking shape in the sixteenth century, were far from obvious. In modern times, understanding of gravitational force has expanded manyfold: gravity is clearly a law throughout the universe—yet some of the more complicated questions regarding gravitational force are far from settled.

HOW IT WORKS

ARISTOTLE'S MODEL

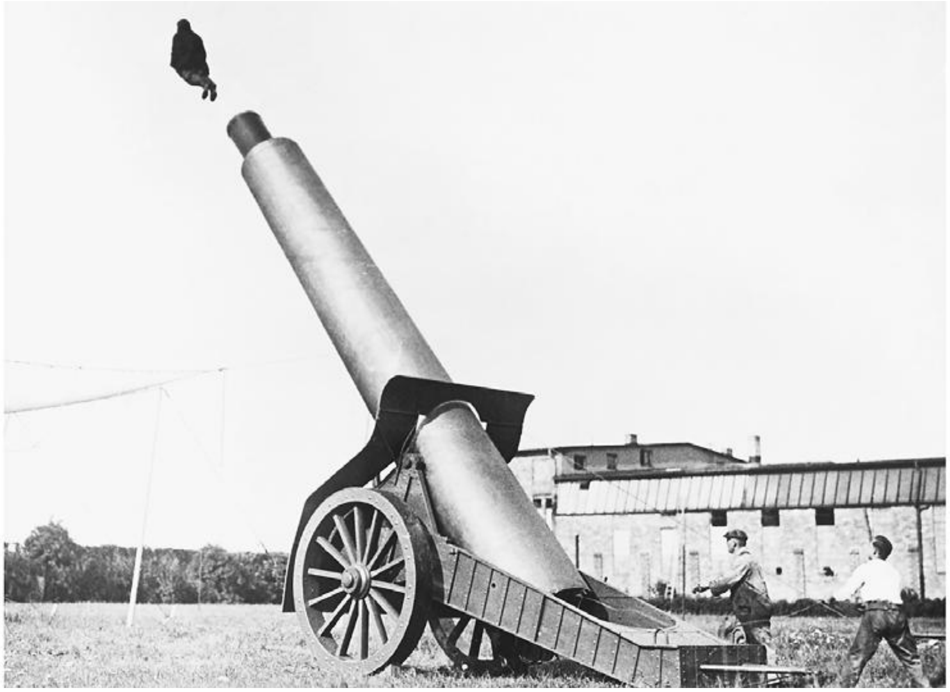
Greek philosophers of the period from the sixth to the fourth century B.C. grappled with a variety of questions concerning the fundamental nature of physical reality, and the forces that bind that reality into a whole. Among the most advanced thinkers of that period was Democritus (c. 460-370 B.C.), who put forth a hypothesis many thousands of years ahead of its time: that all of matter interacts at the atomic level.

Aristotle (384-322 B.C.), however, rejected the explanation offered by Democritus, an unfortunate circumstance given the fact that the great philosopher exerted an incalculable influence on the development of scientific thought. Aristotle's contributions to the advancement of the sciences were many and varied, yet his influence in physics was at least as harmful as it was beneficial. Furthermore, the fact that intellectual progress began slowing after several fruitful centuries of development in Greece only compounded the error. By the time civilization had reached the Middle Ages (c. 500 A.D.) the Aristotelian model of physical reality had been firmly established, and an entire millennium passed before it was successfully challenged.

Wrong though it was in virtually all particulars, the Aristotelian system offered a comforting symmetry amid the troubled centuries of the early medieval period. It must have been reassuring indeed to believe that the physical universe was as simple as the world of human affairs was complex. According to this neat model, all materials on Earth consisted of four elements: earth, water, air, and fire.

Each element had its natural place. Hence, earth was always the lowest, and in some places, earth was covered by water. Water must then be higher, but clearly air was higher still, since it covered earth and water. Highest of all was fire, whose natural place was in the skies above the air. Reflecting these concentric circles were the orbits of the Sun, the Moon, and the five known planets. Their orbital paths, in the Aristotelian model of the universe—a model developed to a great degree by the astronomer Ptolemy (c. 100-170)—were actually spheres that revolved around Earth with clockwork precision.

On Earth, according to the Aristotelian model, objects tended to fall toward the ground in accordance with the admixtures of differing



BECAUSE OF EARTH'S GRAVITY, THE WOMAN BEING SHOT OUT OF THIS CANNON WILL EVENTUALLY FALL TO THE GROUND RATHER THAN ASCEND INTO OUTER SPACE. *(Underwood & Underwood/Corbis. Reproduced by permission.)*

elements they contained. A rock, for instance, was mostly earth, and hence it sought its own level, the lowest of all four elements. For the same reason, a burning fire rose, seeking the heights that were fire's natural domain. It followed from this that an object falls faster or slower, depending on the relative mixtures of elements in it: or, to use more modern terms, the heavier the object, the faster it falls.

GALILEO TAKES UP THE COPERNICAN CHALLENGE

Over the centuries, a small but significant body of scientists and philosophers—each working independent from the other but building on the ideas of his predecessors—slowly began chipping away at the Aristotelian framework. The pivotal challenge came in the early part of the century, and the thinker who put it forward was not a physicist but an astronomer: Nicolaus Copernicus (1473-1543.)

Based on his study of the planets, Copernicus offered an entirely new model of the universe, one that placed the Sun and not Earth at its center. He was not the first to offer such an idea: half a century after Aristotle's death, Aristarchus (fl. 270 B.C.) had a similar idea, but Ptolemy rejected his heliocentric (Sun-centered) model in favor of the geocentric or Earth-centered one. In subsequent centuries, no less a political authority than the Catholic Church gave its approval to the Ptolemaic system. This system seemed to fit well with a literal interpretation of biblical passages concerning God's relationship with man, and man's relationship to the cosmos; hence, the heliocentric model of Copernicus constituted an offense to morality.

For this reason, Copernicus was hesitant to defend his ideas publicly, yet these concepts found their way into the consciousness of European thinkers, causing a paradigm shift so fundamental that it has been dubbed "the Copernican Revolution." Still, Copernicus offered no explanation as to why the planets behaved as they did: hence, the true leader of the Copernican Revolution was not Copernicus himself but Galileo Galilei (1564-1642.)

Initially, Galileo set out to study and defend the ideas of Copernicus through astronomy, but soon the Church forced him to recant. It is said that after issuing a statement in which he refuted the proposition that Earth moves—a direct attack on the static harmony of the Aristotelian/Ptolemaic model—he protested in private: "*E pur si muove!*" (But it does move!) Placed under house arrest by authorities from Rome, he turned his attention to an effort that, ironically, struck the fatal blow against the old model of the cosmos: a proof of the Copernican system according to the laws of physics.

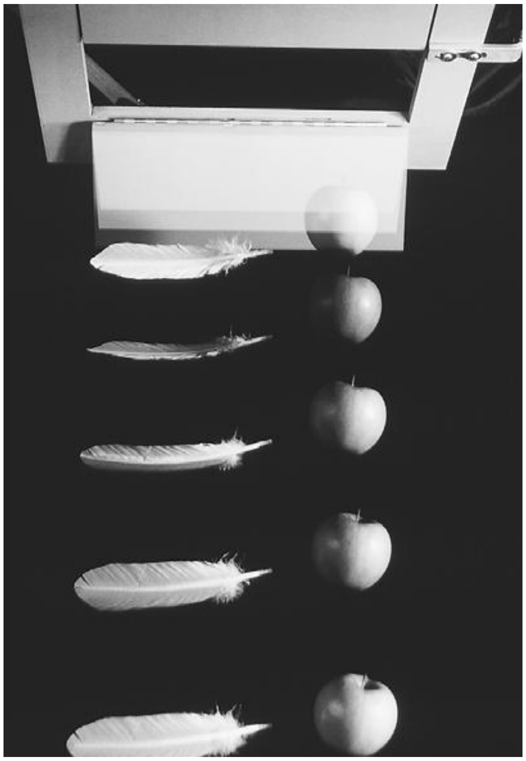
GRAVITATIONAL ACCELERATION. In the process of defending Copernicus, Galileo actually inaugurated the modern history of physics as a science (as opposed to what it had been during the Middle Ages: a nest of suppositions masquerading as knowledge). Specifically, Galileo set out to test the hypothesis that objects fall as they do, not because of their weight, but as a consequence of gravitational force. If this were so, the acceleration of falling bodies would have to be the same, regardless of weight. Of course, it was clear that a stone fell faster than a feather, but Galileo reasoned that this was a result of factors other than weight, and later investigations confirmed that air resistance and friction, not weight, are responsible for this difference.

On the other hand, if one drops two objects that have similar air resistance but differing weight—say, a large stone and a smaller one—they fall at almost exactly the same rate. To test this directly, however, would have been difficult for Galileo: stones fall so fast that, even if dropped from a great height, they would hit the ground too soon for their rate of fall to be tested with the instruments then available.

Instead, Galileo used the motion of a pendulum, and the behavior of objects rolling or sliding down inclined planes, as his models. On the basis of his observations, he concluded that all bodies are subject to a uniform rate of gravitational acceleration, later calibrated at 32 ft (9.8 m) per second. What this means is that for every 32 ft an object falls, it is accelerating at a rate of 32 ft per second as well; hence, after 2 seconds, it falls at the rate of 64 ft (19.6 m) per second; after 3 seconds, at 96 ft (29.4 m) per second, and so on.

Newton Discovers the Principle of Gravity

Building on the work of his distinguished forebear, Sir Isaac Newton (1642-1727)—who, incidentally, was born the same year Galileo died—developed a paradigm for gravitation that, even today, explains the behavior of objects in virtually all situations throughout the universe. Indeed, the Newtonian model reigned until the early twentieth century, when Albert Einstein (1879-1955) challenged it on certain specifics.



THIS PHOTO SHOWS AN APPLE AND A FEATHER BEING DROPPED IN A VACUUM TUBE. BECAUSE OF THE ABSENCE OF AIR RESISTANCE, THE TWO OBJECTS FALL AT THE SAME RATE. *(Photograph by James A. Sugar/Corbis. Reproduced by permission.)*

Even so, Einstein's relativity did not disprove the Newtonian system as Copernicus and Galileo disproved Aristotle's and Ptolemy's theories; rather, it showed the limitations of Newtonian mechanics for describing the behavior of certain objects and phenomena. However, in the ordinary world of day-to-day experience—the world in which stones drop and heavy objects are hard to lift—the Newtonian system still offers the key to how and why things work as they do. This is particularly the case with regard to gravity and gravitation.

Like Galileo, Newton began in part with the aim of testing hypotheses put forth by an astronomer—in this case Johannes Kepler (1571-1630). In the early years of the seventeenth century, Kepler published his three laws of planetary motion, which together identified the elliptical (oval-shaped) path of the planets around the Sun. Kepler had discovered a mathematical relationship that connected the distances of the planets from the Sun to the period of their revolution around it. Like Galileo with Copernicus, Newton sought to generalize these principles to explain, not only how the planets moved, but also why they did.

Almost everyone has heard the story of Newton and the apple—specifically, that while he was sitting under an apple tree, a falling apple struck him on the head, spurring in him a great intuitive leap that led him to form his theory of gravitation. One contemporary biographer, William Stukely, wrote that he and Newton were sitting in a garden under some apple trees when Newton told him that "...he was just in the same situation, as when formerly, the notion of gravitation came into his mind. It was occasion'd by the fall of an apple, as he sat in a contemplative mood. Why should that apple always descend perpendicularly to the ground, he thought to himself. Why should it not go sideways or upwards, but constantly to the earth's centre?"

The tale of Newton and the apple has become a celebrated myth, rather like that of George Washington and the cherry tree. It is an embellishment of actual events: Newton never said that an apple hit him on the head, just that he was thinking about the way that apples fell. Yet the story has become symbolic of the creative intellectual process that occurs when a thinker makes a vast intuitive leap in a matter of moments. Of course, Newton had spent many years contemplating these ideas, and their development required great effort. What is important is that he brought together the best work of his predecessors, yet transcended all that had gone before—and in the process, forged a model that explained a great deal about how the universe functions.

The result was his *Philosophiae Naturalis Principia Mathematica,* or "Mathematical Principles of Natural Philosophy." Published in 1687, the book—usually referred to simply as the *Principia*—was one of the most influential works ever written. In it, Newton presented his three laws of motion, as well as his law of universal gravitation.

The latter stated that every object in the universe attracts every other one with a force proportional to the masses of each, and inversely proportional to the square of the distance between them. This statement requires some clarification with regard to its particulars, after which it will be reintroduced as a mathematical formula.

MASS AND FORCE. The three laws of motion are a subject unto themselves, covered elsewhere in this volume. However, in order to understand gravitation, it is necessary to understand at least a few rudimentary concepts relating to them. The first law identifies inertia as the tendency of an object in motion to remain in motion, and of an object at rest to remain at rest. Inertia is measured by mass, which—as the second law states—is a component of force.

Specifically, the second law of motion states that force is equal to mass multiplied by acceleration. This means that there is an inverse relationship between mass and acceleration: if force remains constant and one of these factors increases, the other must decrease—a situation that will be discussed in some depth below.

Also, as a result of Newton's second law, it is possible to define weight scientifically. People typically assume that mass and weight are the same, and indeed they are on Earth—or at least, they are close enough to be treated as comparable factors. Thus, tables of weights and measures show that a kilogram, the metric unit of mass, is equal to 2.2 pounds, the latter being the principal unit of weight in the British system.

In fact, this is—if not a case of comparing to apples to oranges—certainly an instance of comparing apples to apple pies. In this instance, the kilogram is the "apple" (a fitting Newtonian metaphor!) and the pound the "apple pie." Just as an apple pie contains apples, but other things as well, the pound as a unit of force contains an additional factor, acceleration, not included in the kilo.

BRITISH VS. SI UNITS. Physicists universally prefer the metric system, which is known in the scientific community as SI (an abbreviation of the French *Système International d'Unités*—that is, "International System of Units"). Not only is SI much more convenient to use, due to the fact that it is based on units of 10; but in discussing gravitation, the unequal relationship between kilograms and pounds makes conversion to British units a tedious and ultimately useless task.

Though Americans prefer the British system to SI, and are much more familiar with pounds than with kilos, the British unit of mass—called the slug—is hardly a household word. By contrast, scientists make regular use of the SI unit of force—named, appropriately enough, the newton. In the metric system, a newton (N) is the amount of force required to accelerate 1 kilogram of mass by 1 meter per second squared (m/s^2) Due to the simplicity of using SI over the British system, certain aspects of the discussion below will be presented purely in terms of SI. Where appropriate, however, conversion to British units will be offered.

CALCULATING GRAVITATIONAL FORCE. The law of universal gravitation can be stated as a formula for calculating the gravitational attraction between two objects of a certain mass, m_1 AND M_2: $F_{grav} = G \cdot (m_1M_2)/R^2$. F_{grav} is gravitational force, and r^2 the square of the distance between m_1 and m_2.

As for *G,* in Newton's time the value of this number was unknown. Newton was aware simply that it represented a very small quantity: without it, $(m_1m_2)/r^2$ could be quite sizeable for objects of relatively great mass separated by a relatively small distance. When multiplied by this very small number, however, the gravitational attraction would be revealed to be very small as well. Only in 1798, more than a century after Newton's writing, did English physicist Henry Cavendish (1731-1810) calculate the value of *G.*

As to how Cavendish derived the figure, that is an exceedingly complex subject far beyond the scope of the present discussion. Even to identify *G* as a number is a challenging task. First of all, it is a unit of force multiplied by squared area, then divided by squared mass: in other words, it is expressed in terms of $(N \cdot m^2)/kg^2$, where *N* stands for newtons, *m* for meters, and *kg* for kilograms. Nor is the coefficient, or numerical value, of *G* a whole number such as 1. A figure as large as 1, in fact, is astronomically huge compared to *G,* whose coefficient is $6.67 \cdot 10^{-11}$—in other words, 0.0000000000667.

REAL-LIFE APPLICATIONS

Weight vs. Mass

Before discussing the significance of the gravitational constant, however, at this point it is appropriate to address a few issues that were raised earlier—issues involving mass and weight. In many ways, understanding these properties from the framework of physics requires setting aside everyday notions.

First of all, why the distinction between weight and mass? People are so accustomed to converting pounds to kilos on Earth that the difference is difficult to comprehend, but if one considers the relation of mass and weight in outer space, the distinction becomes much clearer. Mass is the same throughout the universe, making it a much more fundamental characteristic—and hence, physicists typically speak in terms of mass rather than weight.

Weight, on the other hand, differs according to the gravitational pull of the nearest large body. On Earth, a person weighs a certain amount, but on the Moon, this weight is much less, because the Moon possesses less mass than Earth. Therefore, in accordance with Newton's formula for universal gravitation, it exerts less gravitational pull. By contrast, if one were on Jupiter, it would be almost impossible even to stand up, because the pull of gravity on that planet—with its greater mass—would be vastly greater than on Earth.

It should be noted that mass is not at all a function of size: Jupiter does have a greater mass than Earth, but not because it is bigger. Mass, as noted earlier, is purely a measure of inertia: the more resistant an object is to a change in its velocity, the greater its mass. This in itself yields some results that seem difficult to understand as long as one remains wedded to the concept—true enough on Earth—that weight and mass are identical.

A person might weigh less on the Moon, but it would be just as difficult to move that person from a resting position as it would be to do so on Earth. This is because the person's mass, and hence his or her resistance to inertia, has not changed. Again, this is a mentally challenging concept: is not lifting a person, which implies upward acceleration, not an attempt to counteract their inertia when standing still? Does it not follow that their mass has changed? Understanding the distinction requires a greater clarification of the relationship between mass, gravity, and weight.

F = MA. Newton's second law of motion, stated earlier, shows that force is equal to mass multiplied by acceleration, or in shorthand form, $F = ma$. To reiterate a point already made, if one assumes that force is constant, then mass and acceleration must have an inverse relationship. This can be illustrated by performing a simple experiment.

Suppose one were to apply a certain amount of force to an empty shopping cart. Assuming the floor had just enough friction to allow movement, it would be easy for almost anyone to accelerate the shopping cart. Now assume that the shopping cart were filled with heavy lead balls, so that it weighed, say, 1,102 lb (500 kg). If one applied the same force, it would not move.

What has changed, clearly, is the mass of the shopping cart. Because force remained constant, the rate of acceleration would become very small—in this case, almost infinitesimal. In the first case, with an empty shopping cart, the mass was relatively small, so acceleration was relatively high.

Now to return to the subject of lifting someone on the Moon. It is true that in order to lift that person, one would have to overcome inertia, and, in that sense, it would be as difficult as it is on Earth. But the other component of force, acceleration, has diminished greatly.

Weight is, again, a unit of force, but in calculating weight it is useful to make a slight change to the formula $F = ma$. By definition, the acceleration factor in weight is the downward acceleration due to gravity, usually rendered as g. So one's weight is equal to mg—but on the Moon, g is much smaller than it is on Earth, and hence, the same amount of force yields much greater results.

These facts shed new light on a question that bedeviled physicists at least from the time of Aristotle, until Galileo began clarifying the issue some 2,000 years later: why shouldn't an object of greater mass fall at a different rate than one of smaller mass? There are two answers to that question, one general and one specific. The general answer—that Earth exerts more gravitational pull on an object of greater mass—requires a deeper examination of Newton's gravitational formula. But the more specific answer, relating purely to conditions on Earth, is easily addressed by considering the effect of air resistance.

GRAVITY AND AIR RESISTANCE

One of Galileo's many achievements lay in using an idealized model of reality, one that does not take into account the many complex factors that affect the behavior of objects in the real world. This permitted physicists to study processes that apparently defy common sense. For instance, in the real world, an apple does drop at a greater rate of speed than does a feather. However, in a vacuum, they will drop at the same rate. Since Galileo's time, it has become commonplace for physicists to discuss specific processes such as gravity with the assumption that all non-pertinent factors (in this case, air resistance or friction) are nonexistent or irrelevant. This greatly simplified the means of testing hypotheses.

Idealization of reality makes it possible to set aside the things people think they know about the real world, where events are complicated due to friction. The latter may be defined as a force that resists motion when the surface of one object comes into contact with the surface of another. If two balls are released in an environment free from friction—one of them simply dropped while the other is rolled down a curved surface or inclined plane—they will reach the bottom at the same time. This seems to go against everything that is known, but that is only because what people "know" is complicated by variables that have nothing to do with gravity.

The same is true for the behavior of falling objects with regard to air resistance. If air resistance were not a factor, one could fire a cannonball over horizontal space and then, when the ball reached the highest point in its trajectory, release another ball from the same height—and again, they would hit the ground at the same time. This is the case, even though the cannonball that was fired from the cannon has to cover a great deal of horizontal space, whereas the dropped ball does not. The fact is that the rate of acceleration due to gravity will be identical for the two balls, and the fact that the ball fired from a cannon also covers a horizontal distance during that same period is irrelevant.

TERMINAL VELOCITY. In the real world, air resistance creates a powerful drag force on falling objects. The faster the rate of fall, the greater the drag force, until the air resistance forces a leveling in the rate of fall. At this point, the object is said to have reached terminal velocity, meaning that its rate of fall will not increase thereafter. Galileo's idealized model, on the other hand, treated objects as though they were falling in a vacuum—space entirely devoid of matter, including air. In such a situation, the rate of acceleration would continue to grow indefinitely.

By means of a graph, one can compare the behavior of an object falling through air with that of an object falling in a vacuum. If the x axis measures time and the y axis downward speed, the rate of an object falling in a vacuum describes a 60°-angle. In other words, the speed of its descent is increasing at a much faster rate than is the rate of time of its descent—as indeed should be the case, in accordance with gravitational acceleration. The behavior of an object falling through air, on the other hand, describes a curve. Up to a point, the object falls at the same rate as it would in a vacuum, but soon velocity begins to increase at a much slower rate than time. Eventually, the curve levels off at the point where the object experiences terminal velocity.

Air resistance and friction have been mentioned separately as though they were two different forces, but in fact air resistance is simply a prominent form of friction. Hence air resistance exerts an upward force to counter the downward force of mass multiplied by gravity—that is, weight. Since *g* is a constant (32 ft or 9.8 m/sec^2), the greater the weight of the falling object, the longer it takes for air resistance to bring it to terminal velocity.

A feather quickly reaches terminal velocity, whereas it takes much longer for a cannonball to do the same. As a result, a heavier object does take less time to fall, even from a great height, than does a light one—but this is only because of friction, and not because of "elements" seeking their "natural level." Incidentally, if raindrops (which of course fall from a very great height) did not reach terminal velocity, they would cause serious injury by the time they hit the ground.

Applying the Gravitational Formula

Using Newton's gravitational formula, it is relatively easy to calculate the pull of gravity between two objects. It is also easy to see why the attraction is insignificant unless at least one of the objects has enormous mass. In addition, application of the formula makes it clear why *G* (the gravitational constant, as opposed to *g*, the rate of acceleration due to gravity) is such a tiny number.

If two people each have a mass of 45.5 kg (100 lb) and stand 1 m (3.28 ft) apart, m_1m_2 is equal to 2,070 kg (4,555 lb) and r^2 is equal to 1 m^2. Applied to the gravitational formula, this figure is rendered as 2,070 kg^2/1 m^2. This number is then multiplied by gravitational constant, which again is equal to $6.67 \cdot 10^{-11}$ (N • m^2)/kg^2. The result is a net gravitational force of 0.000000138 N (0.00000003 lb)—about the weight of a single-cell organism!

EARTH, GRAVITY, AND WEIGHT. Though it is certainly interesting to calculate the gravitational force between any two people, computations of gravity are only significant for objects of truly great mass. For instance, there is the Earth, which has a mass of $5.98 \cdot 10^{24}$ kg—that is, 5.98 *septillion* (1 followed by 24 zeroes) kilograms. And, of course, Earth's mass is relatively minor compared to that of several planets, not to mention the Sun. Yet Earth exerts enough gravitational pull to keep everything on it—living creatures, manmade structures, mountains and other natural features—stable and in place.

One can calculate Earth's gravitational force on any one person—if one wants to take the time to do so using Newton's formula. In fact, it is much simpler than that: gravitational force is equal to weight, or $m \cdot g$. Thus if a woman weighs 100 lb (445 N), this amount is also equal to the gravitational force exerted on her. By dividing 445 N by the acceleration of gravity—9.8 m/sec^2—it is easy to obtain her mass: 45.4 kg.

The use of the *mg* formula for gravitation helps, once again, to explain why heavier objects do not fall faster than lighter ones. The figure for *g* is a constant, but for the sake of argument, let us assume that it actually becomes larger for objects with a greater mass. This in turn would mean that the gravitational force, or weight, would be bigger than it is—thus creating an irreconcilable logic loop.

Furthermore, one can compare results of two gravitation equations, one measuring the gravitational force between Earth and a large stone, the other measuring the force between Earth and a small stone. (The distance between Earth and each stone is assumed to be the same.) The result will yield a higher quantity for the force exerted on the larger stone—but only because its mass is greater. Clearly, then, the increase of force results only from an increase in mass, not acceleration.

Gravity and Curved Space

As should be clear from Newton's gravitational formula, the force of gravity works both ways: not only does a stone fall toward Earth, but Earth

KEY TERMS

FORCE: The product of mass multiplied by acceleration.

FRICTION: The force that resists motion when the surface of one object comes into contact with the surface of another.

INERTIA: The tendency of an object in motion to remain in motion, and of an object at rest to remain at rest.

INVERSE RELATIONSHIP: A situation involving two variables, in which one of the two increases in direct proportion to the decrease in the other.

LAW OF UNIVERSAL GRAVITATION: A principle, put forth by Sir Isaac Newton (1642-1727), which states that every object in the universe attracts every other one with a force proportional to the masses of each, and inversely proportional to the square of the distance between them.

MASS: A measure of inertia, indicating the resistance of an object to a change in its motion.

TERMINAL VELOCITY: A term describing the rate of fall for an object experiencing the drag force of air resistance. In a vacuum, the object would continue to accelerate with the force of gravity, but in most real-world situations, air resistance creates a powerful drag force that causes a leveling in the object's rate of fall.

VACUUM: Space entirely devoid of matter, including air.

WEIGHT: A measure of the gravitational force on an object; the product of mass multiplied by the acceleration due to gravity.

actually falls toward it. The mass of Earth is so great compared to that of the stone that the movement of Earth is imperceptible—but it does happen. Furthermore, because Earth is round, when one hurls a projectile at a great distance, Earth curves away from the projectile; but eventually gravity itself forces the projectile to the ground.

However, if one were to fire a rocket at 17,700 MPH (28,500 km/h), at every instant of time the projectile is falling toward Earth with the force of gravity—but the curved Earth would be falling away from it at the same moment as well. Hence, the projectile would remain in constant motion around the planet—that is, it would be in orbit.

The same is true of an artificial satellite's orbit around Earth: even as the satellite falls toward Earth, Earth falls away from it. This same relationship exists between Earth and its great natural satellite, the Moon. Likewise, with the Sun and its many satellites, including Earth: Earth plunges toward the Sun with every instant of its movement, but at every instant, the Sun falls away.

WHY IS EARTH ROUND? Note that in the above discussion, it was assumed that Earth and the Sun are round. Everyone knows that to be the case, but why? The answer is "Because they have to be"—that is, gravity will not allow them to be otherwise. In fact, the larger the mass of an object, the greater its tendency toward roundness: specifically, the gravitational pull of its interior forces the surface to assume a relatively uniform shape. There is a relatively small vertical differential for Earth's surface: between the lowest point and the highest point is just 12.28 mi (19.6 km)—not a great distance, considering that Earth's radius is about 4,000 mi (6,400 km).

It is true that Earth bulges near the equator, but this is only because it is spinning rapidly on its axis, and thus responding to the centripetal force of its motion, which produces a centrifugal component. If Earth were standing still, it would be much nearer to the shape of a sphere. On the other hand, an object of less mass is more likely to retain a shape that is far less than spherical. This can be shown by reference to the Martian moons Phobos and Deimos, both of which are oblong—and both of which are tiny, in terms of size and mass, compared to Earth's Moon.

Mars itself has a radius half that of Earth, yet its mass is only about 10% of Earth's. In light of what has been said about mass, shape, and gravity, it should not surprising to learn that Mars is also home to the tallest mountain in the solar system. Standing 15 mi (24 km) high, the volcano Olympus Mons is not only much taller than Earth's tallest peak, Mount Everest (29,028 ft [8,848 m]); it is 22% taller than the distance from the top of Mount Everest to the lowest spot on Earth, the Mariana Trench in the Pacific Ocean (-35,797 ft [-10,911 m])

A spherical object behaves with regard to gravitation as though its mass were concentrated near its center. And indeed, 33% of Earth's mass is at is core (as opposed to the crust or mantle), even though the core accounts for only about 20% of the planet's volume. Geologists believe that the composition of Earth's core must be molten iron, which creates the planet's vast electromagnetic field.

THE FRONTIERS OF GRAVITY. The subject of curvature with regard to gravity can be both a threshold or—as it is here—a point of closure. Investigating questions over perceived anomalies in Newton's description of the behavior of large objects in space led Einstein to his General Theory of Relativity, which posited a curved four-dimensional space-time. This led to entirely new notions concerning gravity, mass, and light. But relativity, as well as its relation to gravity, is another subject entirely. Einstein offered a new understanding of gravity, and indeed of physics itself, that has changed the way thinkers both inside and outside the sciences perceive the universe. Here on Earth, however, gravity behaves much as Newton described it more than three centuries ago.

Meanwhile, research in gravity continues to expand, as a visit to the Web site <www.Gravity.org> reveals. Spurred by studies in relativity, a branch of science called relativistic astrophysics has developed as a synthesis of astronomy and physics that incorporates ideas put forth by Einstein and others. The <www.Gravity.org> site presents studies—most of them too abstruse for a reader who is not a professional scientist—across a broad spectrum of disciplines. Among these is bioscience, a realm in which researchers are investigating the biological effects—such as mineral loss and motion sickness—of exposure to low gravity. The results of such studies will ultimately protect the health of the astronauts who participate in future missions to outer space.

WHERE TO LEARN MORE

Ardley, Neil. *The Science Book of Gravity.* San Diego, CA: Harcourt Brace Jovanovich, 1992.

Beiser, Arthur. *Physics,* 5th ed. Reading, MA: Addison-Wesley, 1991.

Bendick, Jeanne. *Motion and Gravity.* New York: F. Watts, 1972.

Dalton, Cindy Devine. *Gravity.* Vero Beach, FL: Rourke, 2001.

David, Leonard. *"Artificial Gravity and Space Travel." Bio-Science,* March 1992, pp. 155-159.

Exploring Gravity—Curtin University, Australia (Web site). <http://www.curtin.edu.au/curtin/dept/physsci/gravity/> (March 18, 2001).

The Gravity Society (Web site). <http://www.gravity.org> (March 18, 2001).

Nardo, Don. *Gravity: The Universal Force.* San Diego, CA: Lucent Books, 1990.

Rutherford, F. James; Gerald Holton; and Fletcher G. Watson. *Project Physics.* New York: Holt, Rinehart, and Winston, 1981.

Stringer, John. *The Science of Gravity.* Austin, TX: Raintree Steck-Vaughn, 2000.

PROJECTILE MOTION

CONCEPT

A projectile is any object that has been thrown, shot, or launched, and ballistics is the study of projectile motion. Examples of projectiles range from a golf ball in flight, to a curve ball thrown by a baseball pitcher to a rocket fired into space. The flight paths of all projectiles are affected by two factors: gravity and, on Earth at least, air resistance.

HOW IT WORKS

The effects of air resistance on the behavior of projectiles can be quite complex. Because effects due to gravity are much simpler and easier to analyze, and since gravity applies in more situations, we will discuss its role in projectile motion first. In most instances on Earth, of course, a projectile will be subject to both forces, but there may be specific cases in which an artificial vacuum has been created, which means it will only be subjected to the force of gravity. Furthermore, in outer space, gravity—whether from Earth or another body—is likely to be a factor, whereas air resistance (unless or until astronomers find another planet with air) will not be.

The acceleration due to gravity is 32 ft (9.8 m)/sec^2, usually expressed as "per second squared." This means that as every second passes, the speed of a falling object is increasing by 32 ft/sec (9.8 m). Where there is no air resistance, a ball will drop at a velocity of 32 feet per second after one second, 64 ft (19.5 m) per second after two seconds, 96 ft (29.4 m) per second after three seconds, and so on. When an object experiences the ordinary acceleration due to gravity, this figure is rendered in shorthand as *g.* Actually, the figure of 32 ft (9.8 m) per second squared applies at sea level, but since the value of *g* changes little with altitude—it only decreases by 5% at a height of 10 mi (16 km)—it is safe to use this number.

When a plane goes into a high-speed turn, it experiences much higher apparent *g.* This can be as high as 9 *g,* which is almost more than the human body can endure. Incidentally, people call these "*g*-forces," but in fact *g* is not a measure of force but of a single component, acceleration. On the other hand, since force is the product of mass multiplied by acceleration, and since an aircraft subject to a high *g* factor clearly experiences a heavy increase in net force, in that sense, the expression "*g*-force" is not altogether inaccurate.

In a vacuum, where air resistance plays no part, the effects of *g* are clearly demonstrated. Hence a cannonball and a feather, dropped into a vacuum at the same moment, would fall at exactly the same rate and hit bottom at the same time.

The Cannonball or the Feather? Air Resistance vs. Mass

Naturally, air resistance changes the terms of the above equation. As everyone knows, under ordinary conditions, a cannonball falls much faster than a feather, not simply because the feather is lighter than the cannonball, but because the air resists it much better. The speed of descent is a function of air resistance rather than mass, which can be proved with the following experiment. Using two identical pieces of paper—meaning that their mass is exactly the same—wad one up while keeping the other flat. Then drop them. Which one lands first? The wadded piece will fall faster and land first, precisely because it is less air-resistant than the sail-like flat piece.

BECAUSE OF THEIR DESIGN, THE BULLETS IN THIS .357 MAGNUM WILL COME OUT OF THE GUN SPINNING, WHICH GREATLY INCREASES THEIR ACCURACY. *(Photograph by Tim Wright/Corbis. Reproduced by permission.)*

Now to analyze the motion of a projectile in a situation without air resistance. Projectile motion follows the flight path of a parabola, a curve generated by a point moving such that its distance from a fixed point on one axis is equal to its distance from a fixed line on the other axis. In other words, there is a proportional relationship between x and y throughout the trajectory or path of a projectile in motion. Most often this parabola can be visualized as a simple up-and-down curve like the shape of a domed roof. (The Gateway Arch in St. Louis, Missouri, is a steep parabola.)

Instead of referring to the more abstract values of x and y, we will separate projectile motion into horizontal and vertical components. Gravity plays a role only in vertical motion, whereas obviously, horizontal motion is not subject to gravitational force. This means that in the absence of air resistance, the horizontal velocity of a projectile does not change during flight; by contrast, the force of gravity will ultimately reduce its vertical velocity to zero, and this will in turn bring a corresponding drop in its horizontal velocity.

In the case of a cannonball fired at a 45° angle—the angle of maximum efficiency for height and range together—gravity will eventually force the projectile downward, and once it hits the ground, it can no longer continue on its horizontal trajectory. Not, at least, at the same velocity: if you were to thrust a bowling ball forward, throwing it with both hands from the solar plexus, its horizontal velocity would be reduced greatly once gravity forced it to the floor. Nonetheless, the force on the ball would probably be enough (assuming the friction on the floor was not enormous) to keep the ball moving in a horizontal direction for at least a few more feet.

There are several interesting things about the relationship between gravity and horizontal velocity. Assuming, once again, that air resistance is not a factor, the vertical acceleration of a projectile is *g*. This means that when a cannonball is at the highest point of its trajectory, you could simply drop another cannonball from exactly the same height, and they would land at the same moment. This seems counterintuitive, or opposite to common sense: after all, the cannonball that was fired from the cannon has to cover a great deal of horizontal space, whereas the dropped ball does not. Nonetheless, the rate of acceleration due to gravity will be identical for the two balls, and the fact that the ball fired from a cannon also covers a horizontal distance during that same period is purely incidental.

Gravity, combined with the first law of motion, also makes it possible (in theory at least) for a projectile to keep moving indefinitely. This actually does take place at high altitudes, when a satellite is launched into orbit: Earth's gravitational pull, combined with the absence of air resistance or other friction, ensures that the satellite will remain in constant circular motion around the planet. The same is theoretically possible with a cannonball at very low altitudes: if one could fire a ball at 17,700 MPH (28,500 k/mh), the horizontal velocity would be great enough to put the ball into low orbit around Earth's surface.

The addition of air resistance or airflow to the analysis of projectile motion creates a number of complications, including drag, or the force that opposes the forward motion of an object in airflow. Typically, air resistance can create a drag force proportional to the squared value of a projectile's velocity, and this will cause it to fall far short of its theoretical range.

Shape, as noted in the earlier illustration concerning two pieces of paper, also affects air resistance, as does spin. Due to a principle known as the conservation of angular momentum, an object that is spinning tends to keep spinning; moreover, the orientation of the spin axis (the imaginary "pole" around which the object is spinning) tends to remain constant. Thus spin ensures a more stable flight.

REAL-LIFE APPLICATIONS

BULLETS ON A STRAIGHT SPINNING FLIGHT

One of the first things people think of when they hear the word "ballistics" is the study of gunfire patterns for the purposes of crime-solving. Indeed, this application of ballistics is a significant part of police science, because it allows law-enforcement investigators to determine when, where, and how a firearm was used. In a larger sense, however, the term as applied to firearms refers to efforts toward creating a more effective, predictable, and longer bullet trajectory.

From the advent of firearms in the West during the fourteenth century until about 1500, muskets were hopelessly unreliable. This was because the lead balls they fired had not been fitted to the barrel of the musket. When fired, they bounced erratically off the sides of the barrel, and this made their trajectories unpredictable. Compounding this was the unevenness of the lead balls themselves, and this irregularity of shape could lead to even greater irregularities in trajectory.

Around 1500, however, the first true rifles appeared, and these greatly enhanced the accuracy of firearms. The term rifle comes from the "rifling" of the musket barrels: that is, the barrels themselves were engraved with grooves, a process known as rifling. Furthermore, ammunition-makers worked to improve the production process where the musket balls were concerned, producing lead rounds that were more uniform in shape and size.

Despite these improvements, soldiers over the next three centuries still faced many challenges when firing lead balls from rifled barrels. The lead balls themselves, because they were made of a soft material, tended to become misshapen during the loading process. Furthermore, the gunpowder that propelled the lead balls had a tendency to clog the rifle barrel. Most important of all was the fact that these rifles took time to load—and in a situation of battle, this could cost a man his life.

The first significant change came in the 1840s, when in place of lead balls, armies began using bullets. The difference in shape greatly improved the response of rounds to aerodynamic factors. In 1847, Claude-Etienne Minié, a captain in the French army, developed a bullet made of lead, but with a base that was slightly hollow. Thus when fired, the lead in the round tended to expand, filling the barrel's diameter and gripping the rifling.

As a result, the round came out of the barrel end spinning, and continued to spin throughout its flight. Not only were soldiers able to fire their rifles with much greater accuracy, but thanks to the development of chambers and magazines, they could reload more quickly.

CURVE BALLS, DIMPLED GOLF BALLS, AND OTHER TRICKS WITH SPIN

In the case of a bullet, spin increases accuracy, ensuring that the trajectory will follow an expected path. But sometimes spin can be used in more complex ways, as with a curveball thrown by a baseball pitcher.

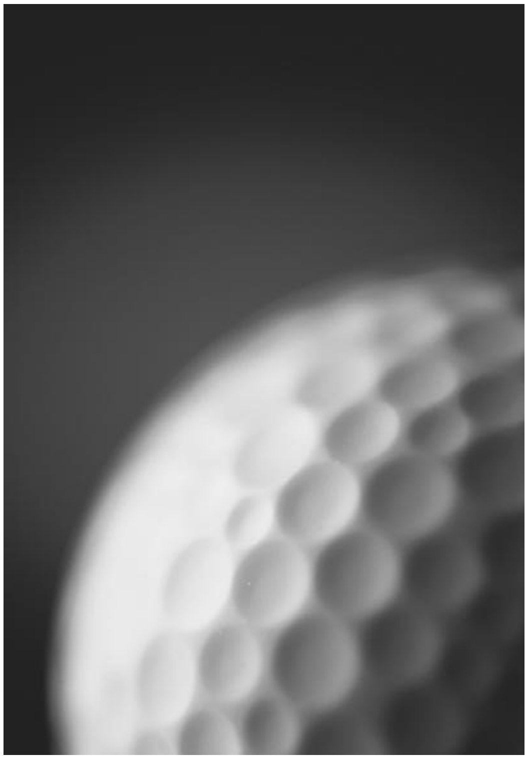
The invention of the curveball is credited to Arthur "Candy" Cummings, who as a pitcher for the Brooklyn Excelsiors at the age of 18 in 1867—an era when baseball was still very young—introduced a new throw he had spent several years perfecting. Snapping as he released the ball, he and the spectators (not to mention the startled batter for the opposing team) watched as the pitch arced, then sailed right past the batter for a strike.

The curveball bedeviled baseball players and fans alike for many years thereafter, and many dismissed it as a type of optical illusion. The debate became so heated that in 1941, both *Life* and *Look* magazines ran features using stop-action photography to show that a curveball truly did curve. Even in 1982, a team of researchers from General Motors (GM) and the Massachusetts Institute of Technology (MIT), working at the behest of *Science* magazine, investigated the curveball to determine if it was more than a mere trick.

In fact, the curveball is a trick, but there is nothing fake about it. As the pitcher releases the ball, he snaps his wrist. This puts a spin on the projectile, and air resistance does the rest. As the ball moves toward the plate, its spin moves against the air, which creates an airstream moving against the trajectory of the ball itself. The airstream splits into two lines, one curving over the ball and one curving under, as the ball sails toward home plate.

For the purposes of clarity, assume that you are viewing the throw from a position between third base and home. Thus, the ball is moving from left to right, and therefore the direction of airflow is from right to left. Meanwhile the ball, as it moves into the airflow, is spinning clockwise. This means that the air flowing over the top of the ball is moving in a direction opposite to the spin, whereas that flowing under it is moving in the same direction as the spin.

This creates an interesting situation, thanks to Bernoulli's principle. The latter, formulated by Swiss mathematician and physicist Daniel Bernoulli (1700-1782), holds that where velocity is high, pressure is low—and vice versa. Bernoulli's principle is of the utmost importance to aerodynamics, and likewise plays a significant role in the operation of a curveball. At the top of the ball, its clockwise spin is moving in a direction opposite to the airflow. This produces drag, slowing the ball, increasing pressure, and thus forcing it downward. At the bottom end of the ball, however, the clockwise motion is flowing with the air, thus resulting in higher velocity and lower pressure. As per Bernoulli's principle, this tends to pull the ball downward.



GOLF BALLS ARE DIMPLED BECAUSE THEY TRAVEL MUCH FARTHER THAN NONDIMPLED ONES. *(Photograph by D. Boone/Corbis. Reproduced by permission.)*

In the 60-ft, 6-in (18.4-m) distance that separates the pitcher's mound from home plate on a regulation major-league baseball field, a curveball can move downward by a foot (0.3048 m) or more. The interesting thing here is that this downward force is almost entirely due to air resistance rather than gravity, though of course gravity eventually brings any pitch to the ground, assuming it has not already been hit, caught, or bounced off a fence.

A curveball represents a case in which spin is used to deceive the batter, but it is just as possible that a pitcher may create havoc at home plate by throwing a ball with little or no spin. This is called a knuckleball, and it is based on the fact that spin in general—though certainly not the deliberate spin of a curveball—tends to ensure a more regular trajectory. Because a knuckleball has no spin, it follows an apparently random path, and thus it can be every bit as tricky for the pitcher as for the batter.

Golf, by contrast, is a sport in which spin is expected: from the moment a golfer hits the ball, it spins backward—and this in turn helps to explain why golf balls are dimpled. Early golf balls, known as featheries, were merely smooth leather pouches containing goose feathers. The smooth surface seemed to produce relatively low drag, and golfers were impressed that a well-hit feathery could travel 150-175 yd (137-160 m).

Then in the late nineteenth century, a professor at St. Andrews University in Scotland realized that a scored or marked ball would travel farther than a smooth one. (The part about St. Andrews may simply be golfing legend, since the course there is regarded as the birthplace of golf in the fifteenth century.) Whatever the case, it is true that a scored ball has a longer trajectory, again as a result of the effect of air resistance on projectile motion.

Airflow produces two varieties of drag on a sphere such as a golf ball: drag due to friction, which is only a small aspect of the total drag, and the much more significant drag that results from the separation of airflow around the ball. As with the curveball discussed earlier, air flows above and below the ball, but the issue here is more complicated than for the curved pitch.

Airflow comes in two basic varieties: laminar, meaning streamlined; or turbulent, indicating an erratic, unpredictable flow. For a jet flying through the air, it is most desirable to create a laminar flow passing over its airfoil, or the curved front surface of the wing. In the case of the golf ball, however, turbulent flow is more desirable.

In laminar flow, the airflow separates quickly, part of it passing over the ball and part passing under. In turbulent flow, however, separation comes later, further back on the ball. Each form of air separation produces a separation region, an area of drag that the ball pulls behind it (so to speak) as it flies through space. But because the separation comes further back on the ball in turbulent flow, the separation region itself is narrower, thus producing less drag.

Clearly, scoring the ball produced turbulent flow, and for a few years in the early twentieth century, manufacturers experimented with designs that included squares, rectangles, and hexagons. In time, they settled on the dimpled design known today. Golf balls made in Britain have 330 dimples, and those in America 336; in either case, the typical drive distance is much, much further than for an unscored ball—180-250 yd (165-229 m).

Powered Projectiles: Rockets and Missiles

The most complex form of projectile widely known in modern life is the rocket or missile. Missiles are unmanned vehicles, most often used in warfare to direct some form of explosive toward an enemy. Rockets, on the other hand, can be manned or unmanned, and may be propulsion vehicles for missiles or for spacecraft. The term rocket can refer either to the engine or to the vehicle it propels.

The first rockets appeared in China during the late medieval period, and were used unsuccessfully by the Chinese against Mongol invaders in the early part of the thirteenth century. Europeans later adopted rocketry for battle, as for instance when French forces under Joan of Arc used crude rockets in an effort to break the siege on Orleans in 1429.

Within a century or so, however, rocketry as a form of military technology became obsolete, though projectile warfare itself remained as effective a method as ever. From the catapults of Roman times to the cannons that appeared in the early Renaissance to the heavy artillery of today, armies have been shooting projectiles against their enemies. The crucial difference between these projectiles and rockets or missiles is that the latter varieties are self-propelled.

Only around the end of World War II did rocketry and missile warfare begin to reappear in new, terrifying forms. Most notable among these was Hitler's V-2 "rocket" (actually a missile), deployed against Great Britain in 1944, but fortunately developed too late to make an impact. The 1950s saw the appearance of nuclear warheads such as the ICBM (intercontinental ballistic missile). These were guided missiles, as opposed to the V-2, which was essentially a huge self-propelled bullet fired toward London.

More effective than the ballistic missile, however, was the cruise missile, which appeared in later decades and which included aerodynamic structures that assisted in guidance and

IN THE CASE OF A ROCKET, LIKE THIS PATRIOT MISSILE BEING LAUNCHED DURING A TEST, PROPULSION COMES BY EXPELLING FLUID—WHICH IN SCIENTIFIC TERMS CAN MEAN A GAS AS WELL AS A LIQUID—FROM ITS REAR END. MOST OFTEN THIS FLUID IS A MASS OF HOT GASES PRODUCED BY A CHEMICAL REACTION INSIDE THE ROCKET'S BODY, AND THIS BACKWARD MOTION CREATES AN EQUAL AND OPPOSITE REACTION FROM THE ATMOSPHERE, PROPELLING THE ROCKET FORWARD. *(Corbis. Reproduced by permission.)*

maneuvering. In addition to guided or unguided, ballistic or aerodynamic, missiles can be classified in terms of source and target: surface-to-surface, air-to-air, and so on. By the 1970s, the United States had developed an extraordinarily sophisticated surface-to-air missile, the Stinger. Stingers proved a decisive factor in the Afghan-Soviet War (1979-89), when U.S.-supplied Afghan guerrillas used them against Soviet aircraft.

In the period from the late 1940s to the late 1980s, the United States, the Soviet Union, and other smaller nuclear powers stockpiled these warheads, which were most effective precisely because they were never used. Thus, U.S. President Ronald Reagan played an important role in ending the Cold War, because his weapons buildup forced the Soviets to spend money they did not have on building their own arsenal. During the aftermath of the Cold War, America and the newly democratized Russian Federation worked to reduce their nuclear stockpiles. Ironically, this was also the period when sophisticated missiles such as the *Patriot* began gaining widespread use in the Persian Gulf War and later conflicts.

Certain properties unite the many varieties of rocket that have existed across time and

KEY TERMS

ACCELERATION: A change in velocity over a given time period.

AERODYNAMIC: Relating to airflow.

BALLISTICS: The study of projectile motion.

DRAG: The force that opposes the forward motion of an object in airflow. In most cases, its opposite is lift.

FIRST LAW OF MOTION: A principle, formulated by Sir Isaac Newton (1642-1727), which states that an object at rest will remain at rest, and an object in motion will remain in motion, at a constant velocity unless or until outside forces act upon it.

FRICTION: Any force that resists the motion of body in relation to another with which it is in contact.

INERTIA: The tendency of an object in motion to remain in motion, and of an object at rest to remain at rest.

LAMINAR: A term describing a streamlined flow, in which all particles move at the same speed and in the same direction. Its opposite is turbulent flow.

LIFT: An aerodynamic force perpendicular to the direction of the wind. In most cases, its opposite is drag.

MASS: A measure of inertia, indicating the resistance of an object to a change in its motion—including a change in velocity.

PARABOLA: A curve generated by a point moving such that its distance from a fixed point on one axis is equal to its distance from a fixed line on the other axis. As a result, between any two points on the parabola there is a proportional relationship between x and y values.

PROJECTILE: Any object that has been thrown, shot, or launched.

SPECIFIC IMPULSE: A measure of rocket fuel efficiency—specifically, the mass that can be lifted by a particular type of fuel for each pound of fuel consumer (that is, the rocket and its contents) per second of operation time. Figures for specific impulse are rendered in seconds.

SPEED: The rate at which the position of an object changes over a given period of time. Unlike velocity, direction is not a component of speed.

THIRD LAW OF MOTION: A principle, which like the first law of motion was formulated by Sir Isaac Newton. The third law states that when one object exerts a force on another, the second object exerts on the first a force equal in magnitude but opposite in direction.

TRAJECTORY: The path of a projectile in motion, a parabola upward and across space.

TURBULENT: A term describing a highly irregular form of flow, in which a fluid is subject to continual changes in speed and direction. Its opposite is laminar flow.

VELOCITY: The speed of an object in a particular direction.

VISCOSITY: The internal friction in a fluid that makes it resistant to flow.

space—including the relatively harmless fireworks used in Fourth of July and New Year's Eve celebrations around the country. One of the key principles that makes rocket propulsion possible is the third law of motion. Sometimes colloquially put as "For every action, there is an equal and opposite reaction," a more scientifically accurate version of this law would be: "When one object exerts a force on another, the second object exerts on the first a force equal in magnitude but opposite in direction."

In the case of a rocket, propulsion comes by expelling fluid—which in scientific terms can mean a gas as well as a liquid—from its rear. Most often this fluid is a mass of hot gases produced by a chemical reaction inside the rocket's body, and this backward motion creates an equal and opposite reaction from the rocket, propelling it forward.

Before it undergoes a chemical reaction, rocket fuel may be either in solid or liquid form inside the rocket's fuel chamber, though it ends up as a gas when expelled. Both solid and liquid varieties have their advantages and disadvantages in terms of safety, convenience, and efficiency in lifting the craft. Scientists calculate efficiency by a number of standards, among them specific impulse, a measure of the mass that can be lifted by a particular type of fuel for each pound of fuel consumed (that is, the rocket and its contents) per second of operation time. Figures for specific impulse are rendered in seconds.

A spacecraft may be divided into segments or stages, which can be released as specific points along the flight in part to increase specific impulse. This was the case with the *Saturn 5* rockets that carried astronauts to the Moon in the period 1969-72, but not with the varieties of space shuttle that have flown regular missions since 1981.

The space shuttle is essentially a hybrid of an airplane and rocket, with a physical structure more like that of an aircraft but with rocket power. In fact, the shuttle uses many rockets to maximize efficiency, utilizing no less than 67 rockets—49 of which run on liquid fuel and the rest on solid fuel—at different stages of its flight.

WHERE TO LEARN MORE

"*Aerodynamics in Sports Equipment.*" *K8AIT Principles of Aeronautics—Advanced* (Web site). <http://muttley.ucdavis.edu/Book/Sports/advanced/index.html> (March 2, 2001).

Asimov, Isaac. *Exploring Outer Space: Rockets, Probes, and Satellites.* Revisions and updates by Francis Reddy. Milwaukee, WI: G. Stevens, 1995.

How in the World? Pleasantville, N.Y.: Reader's Digest, 1990.

"*Interesting Properties of Projectile of Motion*" (Web site). <http://www.phy.ntnu.edu.tw/~hwang/projectile3/projectile3.html> (March 2, 2001).

JBM Small Arms Ballistics (Web site). <http://roadrunner.com/~jbm/index_rgt.html> (March 2, 2001).

The Physics of Projectile Motion (Web site). <http://library.thinkquest.org/2779/> (March 2, 2001).

Richardson, Hazel. *How to Build a Rocket.* Illustrated by Scoular Anderson. New York: F. Watts, 2001.

TORQUE

CONCEPT

Torque is the application of force where there is rotational motion. The most obvious example of torque in action is the operation of a crescent wrench loosening a lug nut, and a close second is a playground seesaw. But torque is also crucial to the operation of gyroscopes for navigation, and of various motors, both internal-combustion and electrical.

HOW IT WORKS

Force, which may be defined as anything that causes an object to move or stop moving, is the linchpin of the three laws of motion formulated by Sir Isaac Newton (1642-1727.) The first law states that an object at rest will remain at rest, and an object in motion will remain in motion, unless or until outside forces act upon it. The second law defines force as the product of mass multiplied by acceleration. According to the third law, when one object exerts a force on another, the second object exerts on the first a force equal in magnitude but opposite in direction.

One way to envision the third law is in terms of an active event—for instance, two balls striking one another. As a result of the impact, each flies backward. Given the fact that the force on each is equal, and that force is the product of mass and acceleration (this is usually rendered with the formula $F = ma$), it is possible to make some predictions regarding the properties of mass and acceleration in this interchange. For instance, if the mass of one ball is relatively small compared to that of the other, its acceleration will be correspondingly greater, and it will thus be thrown backward faster.

On the other hand, the third law can be demonstrated when there is no apparent movement, as for instance, when a person is sitting on a chair, and the chair exerts an equal and opposite force upward. In such a situation, when all the forces acting on an object are in balance, that object is said to be in a state of equilibrium.

Physicists often discuss torque within the context of equilibrium, even though an object experiencing net torque is definitely not in equilibrium. In fact, torque provides a convenient means for testing and measuring the degree of rotational or circular acceleration experienced by an object, just as other means can be used to calculate the amount of linear acceleration. In equilibrium, the net sum of all forces acting on an object should be zero; thus in order to meet the standards of equilibrium, the sum of all torques on the object should also be zero.

REAL-LIFE APPLICATIONS

SEESAWS AND WRENCHES

As for what torque is and how it works, it is best discuss it in relationship to actual objects in the physical world. Two in particular are favorites among physicists discussing torque: a seesaw and a wrench turning a lug nut. Both provide an easy means of illustrating the two ingredients of torque, force and moment arm.

In any object experiencing torque, there is a pivot point, which on the seesaw is the balance-point, and which in the wrench-and-lug nut combination is the lug nut itself. This is the area around which all the forces are directed. In each

A seesaw rotates on and off the ground due to torque imbalance. *(Photograph by Dean Conger/Corbis. Reproduced by permission.)*

case, there is also a place where force is being applied. On the seesaw, it is the seats, each holding a child of differing weight. In the realm of physics, weight is actually a variety of force.

Whereas force is equal to mass multiplied by acceleration, weight is equal to mass multiplied by the acceleration due to gravity. The latter is equal to 32 ft (9.8 m)/sec^2. This means that for every second that an object experiencing gravitational force continues to fall, its velocity increases at the rate of 32 ft or 9.8 m per second. Thus, the formula for weight is essentially the same as that for force, with a more specific variety of acceleration substituted for the generalized term in the equation for force.

As for moment arm, this is the distance from the pivot point to the vector on which force is being applied. Moment arm is always perpendicular to the direction of force. Consider a wrench operating on a lug nut. The nut, as noted earlier, is the pivot point, and the moment arm is the distance from the lug nut to the place where the person operating the wrench has applied force. The torque that the lug nut experiences is the product of moment arm multiplied by force.

In English units, torque is measured in pound-feet, whereas the metric unit is Newton-meters, or N•m. (One newton is the amount of force that, when applied to 1 kg of mass, will give it an acceleration of 1 m/sec^2). Hence if a person were to a grip a wrench 9 in (23 cm) from the pivot point, the moment arm would be 0.75 ft (0.23 m.) If the person then applied 50 lb (11.24 N) of force, the lug nut would be experiencing 37.5 pound-feet (2.59 N•m) of torque.

The greater the amount of torque, the greater the tendency of the object to be put into rotation. In the case of a seesaw, its overall design, in particular the fact that it sits on the ground, means that its board can never undergo anything close to 360° rotation; nonetheless, the board does rotate within relatively narrow parameters. The effects of torque can be illustrated by imagining the clockwise rotational behavior of a seesaw viewed from the side, with a child sitting on the left and a teenager on the right.

Suppose the child weighs 50 lb (11.24 N) and sits 3 ft (0.91 m) from the pivot point, giving her side of the seesaw a torque of 150 pound-feet (10.28 N•m). On the other side, her teenage sister weighs 100 lb (22.48 N) and sits 6 ft (1.82 m) from the center, creating a torque of 600 pound-feet (40.91 N•m). As a result of the torque imbalance, the side holding the teenager will rotate clockwise, toward the ground, causing the child's side to also rotate clockwise—off the ground.

TORQUE, ALONG WITH ANGULAR MOMENTUM, IS THE LEADING FACTOR DICTATING THE MOTION OF A GYROSCOPE. HERE, A WOMAN RIDES INSIDE A GIANT GYROSCOPE AT AN AMUSEMENT PARK. *(Photograph by Richard Cummins/Corbis. Reproduced by permission.)*

In order for the two to balance one another perfectly, the torque on each side has to be adjusted. One way would be by changing weight, but a more likely remedy is a change in position, and therefore, of moment arm. Since the teenager weighs exactly twice as much as the child, the moment arm on the child's side must be exactly twice as long as that on the teenager's.

Hence, a remedy would be for the two to switch positions with regard to the pivot point. The child would then move out an additional 3 ft (.91 m), to a distance of 6 ft (1.83 m) from the pivot, and the teenager would cut her distance from the pivot point in half, to just 3 ft (.91 m). In fact, however, any solution that gave the child a moment arm twice as long as that of the teenager would work: hence, if the teenager sat 1 ft (.3 m) from the pivot point, the child should be at 2 ft (.61 m) in order to maintain the balance, and so on.

On the other hand, there are many situations in which you may be unable to increase force, but can increase moment arm. Suppose you were trying to disengage a particularly stubborn lug nut, and after applying all your force, it still would not come loose. The solution would be to increase moment arm, either by grasping the wrench further from the pivot point, or by using a longer wrench.

For the same reason, on a door, the knob is placed as far as possible from the hinges. Here the hinge is the pivot point, and the door itself is the moment arm. In some situations of torque, however, moment arm may extend over "empty space," and for this reason, the handle of a wrench is not exactly the same as its moment arm. If one applies force on the wrench at a 90°-angle to the handle, then indeed handle and moment arm are identical; however, if that force were at a 45° angle, then the moment arm would be outside the handle, because moment arm and force are always perpendicular. And if one were to pull the wrench away from the lug nut, then there would be 0° difference between the direction of force and the pivot point—meaning that moment arm (and hence torque) would also be equal to zero.

Gyroscopes

A gyroscope consists of a wheel-like disk, called a flywheel, mounted on an axle, which in turn is mounted on a larger ring perpendicular to the plane of the wheel itself. An outer circle on the same plane as the flywheel provides structural stability, and indeed, the gyroscope may include several such concentric rings. Its focal point, however, is the flywheel and the axle. One end of the axle is typically attached to some outside object, while the other end is left free to float.

Once the flywheel is set spinning, gravity has a tendency to pull the unattached end of the axle downward, rotating it on an axis perpendicular to that of the flywheel. This should cause the gyroscope to fall over, but instead it begins to spin a third axis, a horizontal axis perpendicular both to the plane of the flywheel and to the direction of gravity. Thus, it is spinning on three axes, and as a result becomes very stable—that is, very resistant toward outside attempts to upset its balance.

This in turn makes the gyroscope a valued instrument for navigation: due to its high degree of gyroscopic inertia, it resists changes in orientation, and thus can guide a ship toward its destination. Gyroscopes, rather than magnets, are often the key element in a compass. A magnet will point to magnetic north, some distance from "true north" (that is, the North Pole.) But with a gyroscope whose axle has been aligned with true north before the flywheel is set spinning, it is possible to possess a much more accurate directional indicator. For this reason, gyroscopes are used on airplanes—particularly those flying over the poles—as well as submarines and even the Space Shuttle.

Torque, along with angular momentum, is the leading factor dictating the motion of a gyroscope. Think of angular momentum as the momentum (mass multiplied by velocity) that a turning object acquires. Due to a principle known as the conservation of angular momentum, a spinning object has a tendency to reach a constant level of angular momentum, and in order to do this, the sum of the external torques acting on the system must be reduced to zero. Thus angular momentum "wants" or "needs" to cancel out torque.

The "right-hand rule" can help you to understand the torque in a system such as the gyroscope. If you extend your right hand, palm downward, your fingers are analogous to the moment arm. Now if you curl your fingers downward, toward the ground, then your fingertips point in the direction of *g*—that is, gravitational force. At that point, your thumb (involuntarily, due to the bone structure of the hand) points in the direction of the torque vector.

When the gyroscope starts to spin, the vectors of angular momentum and torque are at odds with one another. Were this situation to persist, it would destabilize the gyroscope; instead, however, the two come into alignment. Using the right-hand rule, the torque vector on a gyroscope is horizontal in direction, and the vector of angular momentum eventually aligns with

KEY TERMS

ACCELERATION: A change in velocity over a given time period.

EQUILIBRIUM: A situation in which the forces acting upon an object are in balance.

FORCE: The product of mass multiplied by acceleration.

INERTIA: The tendency of an object in motion to remain in motion, and of an object at rest to remain at rest.

MASS: A measure of inertia, indicating the resistance of an object to a change in its motion—including a change in velocity.

MOMENT ARM: For an object experiencing torque, moment arm is the distance from the pivot or balance point to the vector on which force is being applied. Moment arm is always perpendicular to the direction of force.

SPEED: The rate at which the position of an object changes over a given period of time.

TORQUE: The product of moment arm multiplied by force.

VECTOR: A quantity that possesses both magnitude and direction. By contrast, a scalar quantity is one that possesses only magnitude, with no specific direction.

VELOCITY: The speed of an object in a particular direction.

WEIGHT: A measure of the gravitational force on an object; the product of mass multiplied by the acceleration due to gravity.

it. To achieve this, the gyroscope experiences what is known as gyroscopic precession, pivoting along its support post in an effort to bring angular momentum into alignment with torque. Once this happens, there is no net torque on the system, and the conservation of angular momentum is in effect.

Torque in Complex Machines

Torque is a factor in several complex machines such as the electric motor that—with variations—runs most household appliances. It is especially important to the operation of automobiles, playing a significant role in the engine and transmission.

An automobile engine produces energy, which the pistons or rotor convert into torque for transmission to the wheels. Though torque is greatest at high speeds, the amount of torque needed to operate a car does not always vary proportionately with speed. At moderate speeds and on level roads, the engine does not need to provide a great deal of torque. But when the car is starting, or climbing a steep hill, it is important that the engine supply enough torque to keep the car running; otherwise it will stall. To allocate torque and speed appropriately, the engine may decrease or increase the number of revolutions per minute to which the rotors are subjected.

Torque comes from the engine, but it has to be supplied to the transmission. In an automatic transmission, there are two principal components: the automatic gearbox and the torque converter. It is the job of the torque converter to transmit power from the flywheel of the engine to the gearbox, and it has to do so as smoothly as possible. The torque converter consists of three elements: an impeller, which is turned by the engine flywheel; a reactor that passes this motion on to a turbine; and the turbine itself, which turns the input shaft on the automatic gearbox. An infusion of oil to the converter assists the impeller and turbine in synchronizing movement, and this alignment of elements in the torque converter creates a smooth relationship between engine and gearbox. This also leads to an increase in the car's overall torque—that is, its turning force.

Torque is also important in the operation of electric motors, found in everything from vacuum cleaners and dishwashers to computer printers and videocassette recorders to subway systems and water-pumping stations. Torque in the context of electricity involves reference to a number of concepts beyond the scope of this discussion: current, conduction, magnetic field, and other topics relevant to electromagnetic force.

WHERE TO LEARN MORE

Beiser, Arthur. *Physics,* 5th ed. Reading, MA: Addison-Wesley, 1991.

Macaulay, David. *The New Way Things Work.* Boston: Houghton Mifflin, 1998.

"*Rotational Motion.*" *Physics Department, University of Guelph* (Web site). <http://www.physics.uoguelph.ca/tutorials/torque/> (March 4, 2001).

"*Rotational Motion—Torque.*" *Lee College* (Web site). <http://www.lee.edu/mathscience/physics/physics/Courses/LabManual/2b/2b.html> (March 4, 2001).

Schweiger, Peggy E. "*Torque*" (Web site). <http://www.cyberclassrooms.net/~pschweiger/rotmot.html> (March 4, 2001).

"*Torque and Rotational Motion*" (Web site). <http://online.cctt.org/curriculumguide/units/torque.asp> (March 4, 2001).

FLUID MECHANICS

FLUID MECHANICS

CONCEPT

The term "fluid" in everyday language typically refers only to liquids, but in the realm of physics, fluid describes any gas or liquid that conforms to the shape of its container. Fluid mechanics is the study of gases and liquids at rest and in motion. This area of physics is divided into fluid statics, the study of the behavior of stationary fluids, and fluid dynamics, the study of the behavior of moving, or flowing, fluids. Fluid dynamics is further divided into hydrodynamics, or the study of water flow, and aerodynamics, the study of airflow. Applications of fluid mechanics include a variety of machines, ranging from the waterwheel to the airplane. In addition, the study of fluids provides an understanding of a number of everyday phenomena, such as why an open window and door together create a draft in a room.

HOW IT WORKS

The Contrast Between Fluids and Solids

To understand fluids, it is best to begin by contrasting their behavior with that of solids. Whereas solids possess a definite volume and a definite shape, these physical characteristics are not so clearly defined for fluids. Liquids, though they possess a definite volume, have no definite shape—a factor noted above as one of the defining characteristics of fluids. As for gases, they have neither a definite shape nor a definite volume.

One of several factors that distinguishes fluids from solids is their response to compression, or the application of pressure in such a way as to reduce the size or volume of an object. A solid is highly noncompressible, meaning that it resists compression, and if compressed with a sufficient force, its mechanical properties alter significantly. For example, if one places a drinking glass in a vise, it will resist a small amount of pressure, but a slight increase will cause the glass to break.

Fluids vary with regard to compressibility, depending on whether the fluid in question is a liquid or a gas. Most gases tend to be highly compressible—though air, at low speeds at least, is not among them. Thus, gases such as propane fuel can be placed under high pressure. Liquids tend to be noncompressible: unlike a gas, a liquid can be compressed significantly, yet its response to compression is quite different from that of a solid—a fact illustrated below in the discussion of hydraulic presses.

One way to describe a fluid is "anything that flows"—a behavior explained in large part by the interaction of molecules in fluids. If the surface of a solid is disturbed, it will resist, and if the force of the disturbance is sufficiently strong, it will deform—as for instance, when a steel plate begins to bend under pressure. This deformation will be permanent if the force is powerful enough, as was the case in the above example of the glass in a vise. By contrast, when the surface of a liquid is disturbed, it tends to flow.

MOLECULAR BEHAVIOR OF FLUIDS AND SOLIDS. At the molecular level, particles of solids tend to be definite in their arrangement and close to one another. In the case of liquids, molecules are close in proximity, though not as much so as solid molecules, and the arrangement is random. Thus, with a glass of water, the molecules of glass (which at

In a wide, unconstricted region, a river flows slowly. However, if its flow is narrowed by canyon walls, as with Wyoming's Bighorn River, then it speeds up dramatically. *(Photograph by Kevin R. Morris/Corbis. Reproduced by permission.)*

relatively low temperatures is a solid) in the container are fixed in place while the molecules of water contained by the glass are not. If one portion of the glass were moved to another place on the glass, this would change its structure. On the other hand, no significant alteration occurs in the character of the water if one portion of it is moved to another place within the entire volume of water in the glass.

As for gas molecules, these are both random in arrangement and far removed in proximity. Whereas solid particles are slow-moving and have a strong attraction to one another, liquid molecules move at moderate speeds and exert a moderate attraction on each other. Gas molecules are extremely fast-moving and exert little or no attraction.

Thus, if a solid is released from a container pointed downward, so that the force of gravity moves it, it will fall as one piece. Upon hitting a floor or other surface, it will either rebound, come to a stop, or deform permanently. A liquid, on the other hand, will disperse in response to impact, its force determining the area over which the total volume of liquid is distributed. But for a gas, assuming it is lighter than air, the downward pull of gravity is not even required to disperse it: once the top on a container of gas is released, the molecules begin to float outward.

Fluids Under Pressure

As suggested earlier, the response of fluids to pressure is one of the most significant aspects of fluid behavior and plays an important role within both the statics and dynamics subdisciplines of fluid mechanics. A number of interesting principles describe the response to pressure, on the part of both fluids at rest inside a container, and fluids which are in a state of flow.

Within the realm of hydrostatics, among the most important of all statements describing the behavior of fluids is Pascal's principle. This law is named after Blaise Pascal (1623-1662), a French mathematician and physicist who discovered that the external pressure applied on a fluid is transmitted uniformly throughout its entire body. The understanding offered by Pascal's principle later became the basis for one of the most important machines ever developed, the hydraulic press.

HYDROSTATIC PRESSURE AND BUOYANCY. Some nineteen centuries before Pascal, the Greek mathematician, physicist, and inventor Archimedes (c. 287-212 B.C.) discovered a precept of fluid statics that had implications at least as great as those of Pascal's principle. This was Archimedes's principle, which explains the buoyancy of an object immersed in fluid. According to Archimedes's principle, the buoyant force exerted on the object is equal to the weight of the fluid it displaces.

Buoyancy explains both how a ship floats on water, and how a balloon floats in the air. The pressures of water at the bottom of the ocean, and of air at the surface of Earth, are both examples of hydrostatic pressure—the pressure that exists at any place in a body of fluid due to the weight of the fluid above. In the case of air pressure, air is pulled downward by the force of Earth's gravitation, and air along the planet's surface has greater pressure due to the weight of the air above it. At great heights above Earth's surface, however, the gravitational force is diminished, and thus the air pressure is much smaller.

Water, too, is pulled downward by gravity, and as with air, the fluid at the bottom of the ocean has much greater pressure due to the weight of the fluid above it. Of course, water is much heavier than air, and therefore, water at even a moderate depth in the ocean has enormous pressure. This pressure, in turn, creates a buoyant force that pushes upward.

If an object immersed in fluid—a balloon in the air, or a ship on the ocean—weighs less that the fluid it displaces, it will float. If it weighs more, it will sink or fall. The balloon itself may be "heavier than air," but it is not as heavy as the air it has displaced. Similarly, an aircraft carrier contains a vast weight in steel and other material, yet it floats, because its weight is not as great as that of the displaced water.

BERNOULLI'S PRINCIPLE. Archimedes and Pascal contributed greatly to what became known as fluid statics, but the father of fluid mechanics, as a larger realm of study, was the Swiss mathematician and physicist Daniel Bernoulli (1700-1782). While conducting experiments with liquids, Bernoulli observed that when the diameter of a pipe is reduced, the water flows faster. This suggested to him that some force must be acting upon the water, a force that he reasoned must arise from differences in pressure.

Specifically, the slower-moving fluid in the wider area of pipe had a greater pressure than the portion of the fluid moving through the narrower part of the pipe. As a result, he concluded that pressure and velocity are inversely related—in other words, as one increases, the other decreases. Hence, he formulated Bernoulli's principle, which states that for all changes in movement, the sum of static and dynamic pressure in a fluid remains the same.

A fluid at rest exerts pressure—what Bernoulli called "static pressure"—on its container. As the fluid begins to move, however, a portion of the static pressure—proportional to the speed of the fluid—is converted to what Bernoulli called dynamic pressure, or the pressure of movement. In a cylindrical pipe, static pressure is exerted perpendicular to the surface of the container, whereas dynamic pressure is parallel to it.

According to Bernoulli's principle, the greater the velocity of flow in a fluid, the greater the dynamic pressure and the less the static pressure. In other words, slower-moving fluid exerts greater pressure than faster-moving fluid. The discovery of this principle ultimately made possible the development of the airplane.

REAL-LIFE APPLICATIONS

Bernoulli's Principle in Action

As fluid moves from a wider pipe to a narrower one, the volume of the fluid that moves a given distance in a given time period does not change. But since the width of the narrower pipe is smaller, the fluid must move faster (that is, with greater dynamic pressure) in order to move the same amount of fluid the same distance in the same amount of time. Observe the behavior of a river: in a wide, unconstricted region, it flows slowly, but if its flow is narrowed by canyon walls, it speeds up dramatically.

Bernoulli's principle ultimately became the basis for the airfoil, the design of an airplane's wing when seen from the end. An airfoil is shaped like an asymmetrical teardrop laid on its side, with the "fat" end toward the airflow. As air hits the front of the airfoil, the airstream divides, part of it passing over the wing and part passing under. The upper surface of the airfoil is curved, however, whereas the lower surface is much straighter.

As a result, the air flowing over the top has a greater distance to cover than the air flowing under the wing. Since fluids have a tendency to compensate for all objects with which they come into contact, the air at the top will flow faster to meet the other portion of the airstream, the air flowing past the bottom of the wing, when both reach the rear end of the airfoil. Faster airflow, as demonstrated by Bernoulli, indicates lower pressure, meaning that the pressure on the bottom of the wing keeps the airplane aloft.

CREATING A DRAFT. Among the most famous applications of Bernoulli's principle is its use in aerodynamics, and this is discussed in the context of aerodynamics itself elsewhere in this book. Likewise, a number of other applications of Bernoulli's principle are examined in an essay devoted to that topic. Bernoulli's principle, for instance, explains why a shower curtain tends to billow inward when the water is turned on; in addition, it shows why an open window and door together create a draft.

Suppose one is in a hotel room where the heat is on too high, and there is no way to adjust the thermostat. Outside, however, the air is cold, and thus, by opening a window, one can presumably cool down the room. But if one opens the window without opening the front door of the room, there will be little temperature change. The only way to cool off will be by standing next to the window: elsewhere in the room, the air will be every bit as stuffy as before. But if the door leading to the hotel hallway is opened, a nice cool breeze will blow through the room. Why?

With the door closed, the room constitutes an area of relatively high pressure compared to the pressure of the air outside the window. Because air is a fluid, it will tend to flow into the room, but once the pressure inside reaches a certain point, it will prevent additional air from entering. The tendency of fluids is to move from high-pressure to low-pressure areas, not the other way around. As soon as the door is opened, the relatively high-pressure air of the room flows into the relatively low-pressure area of the hallway. As a result, the air pressure in the room is reduced, and the air from outside can now enter. Soon a wind will begin to blow through the room.

A WIND TUNNEL. The above scenario of wind flowing through a room describes a rudimentary wind tunnel. A wind tunnel is a chamber built for the purpose of examining the characteristics of airflow in contact with solid objects, such as aircraft and automobiles. The wind tunnel represents a safe and judicious use of the properties of fluid mechanics. Its purpose is to test the interaction of airflow and solids in relative motion: in other words, either the aircraft has to be moving against the airflow, as it does in flight, or the airflow can be moving against a stationary aircraft. The first of these choices, of course, poses a number of dangers; on the other hand, there is little danger in exposing a stationary craft to winds at speeds simulating that of the aircraft in flight.

The first wind tunnel was built in England in 1871, and years later, aircraft pioneers Orville (1871-1948) and Wilbur (1867-1912) Wright used a wind tunnel to improve their planes. By the late 1930s, the U.S. National Advisory Committee for Aeronautics (NACA) was building wind tunnels capable of creating speeds equal to 300 MPH (480 km/h); but wind tunnels built after World War II made these look primitive. With the development of jet-powered flight, it became necessary to build wind tunnels capable of simulating winds at the speed of sound—760 MPH (340 m/s). By the 1950s, wind tunnels were being used to simulate hypersonic speeds—that is, speeds of Mach 5 (five times the speed of sound) and above. Researchers today use helium to create wind blasts at speeds up to Mach 50.

Fluid Mechanics for Performing Work

HYDRAULIC PRESSES. Though applications of Bernoulli's principle are among the most dramatic examples of fluid mechanics in operation, the everyday world is filled with instances of other ideas at work. Pascal's principle, for instance, can be seen in the operation of any number of machines that represent variations on the idea of a hydraulic press. Among these is the hydraulic jack used to raise a car off the floor of an auto mechanic's shop.

Beneath the floor of the shop is a chamber containing a quantity of fluid, and at either end of the chamber are two large cylinders side by side. Each cylinder holds a piston, and valves control flow between the two cylinders through the channel of fluid that connects them. In accordance with Pascal's principle, when one applies force by pressing down the piston in one cylinder (the input cylinder), this yields a uniform pressure that causes output in the second cylinder, pushing up a piston that raises the car.

Another example of a hydraulic press is the hydraulic ram, which can be found in machines ranging from bulldozers to the hydraulic lifts used by firefighters and utility workers to reach heights. In a hydraulic ram, however, the characteristics of the input and output cylinders are reversed from those of a car jack. For the car jack, the input cylinder is long and narrow, while the output cylinder is wide and short. This is because the purpose of a car jack is to raise a heavy object through a relatively short vertical range of movement—just high enough so that the mechanic can stand comfortably underneath the car.

In the hydraulic ram, the input or master cylinder is short and squat, while the output or slave cylinder is tall and narrow. This is because the hydraulic ram, in contrast to the car jack, carries a much lighter cargo (usually just one person) through a much greater vertical range—for instance, to the top of a tree or building.

PUMPS. A pump is a device made for moving fluid, and it does so by utilizing a pressure difference, causing the fluid to move from an area of higher pressure to one of lower pressure. Its operation is based on aspects both of Pascal's and Bernoulli's principles—though, of course, humans were using pumps thousands of years before either man was born.

A siphon hose used to draw gas from a car's fuel tank is a very simple pump. Sucking on one end of the hose creates an area of low pressure compared to the relatively high-pressure area of the gas tank. Eventually, the gasoline will come out of the low-pressure end of the hose.

The piston pump, slightly more complex, consists of a vertical cylinder along which a piston rises and falls. Near the bottom of the cylinder are two valves, an inlet valve through which fluid flows into the cylinder, and an outlet valve through which fluid flows out. As the piston moves upward, the inlet valve opens and allows fluid to enter the cylinder. On the downstroke, the inlet valve closes while the outlet valve opens, and the pressure provided by the piston forces the fluid through the outlet valve.

One of the most obvious applications of the piston pump is in the engine of an automobile. In this case, of course, the fluid being pumped is gasoline, which pushes the pistons up and down by providing a series of controlled explosions created by the spark plug's ignition of the gas. In another variety of piston pump—the kind used to inflate a basketball or a bicycle tire—air is the fluid being pumped. Then there is a pump for water. Pumps for drawing usable water from the ground are undoubtedly the oldest known variety, but there are also pumps designed to remove water from areas where it is undesirable; for example, a bilge pump, for removing water from a boat, or the sump pump used to pump flood water out of a basement.

PUMPS FOR DRAWING USABLE WATER FROM THE GROUND ARE UNDOUBTEDLY THE OLDEST PUMPS KNOWN. *(Photograph by Richard Cummins/Corbis. Reproduced by permission.)*

FLUID POWER. For several thousand years, humans have used fluids—in particular water—to power a number of devices. One of the great engineering achievements of ancient times was the development of the waterwheel, which included a series of buckets along the rim that made it possible to raise water from the river below and disperse it to other points. By about 70 B.C., Roman engineers recognized that they could use the power of water itself to turn wheels and grind grain. Thus, the waterwheel became one of the first mechanisms in which an inanimate

KEY TERMS

AERODYNAMICS: An area of fluid dynamics devoted to studying the properties and characteristics of airflow.

ARCHIMEDES'S PRINCIPLE: A rule of physics stating that the buoyant force of an object immersed in fluid is equal to the weight of the fluid displaced by the object. It is named after the Greek mathematician, physicist, and inventor, Archimedes (c. 287-212 B.C.), who first identified it.

BERNOULLI'S PRINCIPLE: A proposition, credited to Swiss mathematician and physicist Daniel Bernoulli (1700-1782), which maintains that slower-moving fluid exerts greater pressure than faster-moving fluid.

BUOYANCY: The tendency of an object immersed in a fluid to float. This can be explained by Archimedes's principle.

COMPRESSION: To reduce in size or volume by applying pressure.

FLUID: Any substance, whether gas or liquid, that conforms to the shape of its container.

FLUID DYNAMICS: An area of fluid mechanics devoted to studying of the behavior of moving, or flowing, fluids. Fluid dynamics is further divided into hydrodynamics and aerodynamics.

FLUID MECHANICS: The study of the behavior of gases and liquids at rest and in motion. The major divisions of fluid mechanics are fluid statics and fluid dynamics.

FLUID STATICS: An area of fluid mechanics devoted to studying the behavior of stationary fluids.

HYDRODYNAMICS: An area of fluid dynamics devoted to studying the properties and characteristics of water flow.

HYDROSTATIC PRESSURE: The pressure that exists at any place in a body of fluid due to the weight of the fluid above.

PASCAL'S PRINCIPLE: A statement, formulated by French mathematician and physicist Blaise Pascal (1623-1662), which holds that the external pressure applied on a fluid is transmitted uniformly throughout the entire body of that fluid.

PRESSURE: The ratio of force to surface area, when force is applied in a direction perpendicular to that surface.

TURBINE: A machine that converts the kinetic energy (the energy of movement) in fluids to useable mechanical energy by passing the stream of fluid through a series of fixed and moving fans or blades.

WIND TUNNEL: A chamber built for the purpose of examining the characteristics of airflow in relative motion against solid objects such as aircraft and automobiles.

source (as opposed to the effort of humans or animals) created power.

The water clock, too, was another ingenious use of water developed by the ancients. It did not use water for power; rather, it relied on gravity—a concept only dimly understood by ancient peoples—to move water from one chamber of the clock to another, thus, marking a specific interval of time. The earliest clocks were sundials, which were effective for measuring time, provided the Sun was shining, but which were less useful for measuring periods shorter than an hour. Hence, the development of the hourglass, which used sand, a solid that in larger quantities exhibits the behavior of a fluid. Then, in about 270 B.C., Ctesibius of Alexandria (fl. c. 270-250 B.C.) used gearwheel technology to devise a constant-flow water clock called a "clepsydra." Use of water clocks prevailed for more than a thousand years, until the advent of the first mechanical clocks.

During the medieval period, fluids provided power to windmills and water mills, and at the dawn of the Industrial Age, engineers began applying fluid principles to a number of sophisticated machines. Among these was the turbine, a machine that converts the kinetic energy (the energy of movement) in fluids to useable mechanical energy by passing the stream of fluid through a series of fixed and moving fans or blades. A common house fan is an example of a turbine in reverse: the fan adds energy to the passing fluid (air), whereas a turbine extracts energy from fluids such as air and water.

The turbine was developed in the mid-eighteenth century, and later it was applied to the extraction of power from hydroelectric dams, the first of which was constructed in 1894. Today, hydroelectric dams provide electric power to millions of homes around the world. Among the most dramatic examples of fluid mechanics in action, hydroelectric dams are vast in size and equally impressive in the power they can generate using a completely renewable resource: water.

A hydroelectric dam forms a huge steel-and-concrete curtain that holds back millions of tons of water from a river or other body. The water nearest the top—the "head" of the dam—has enormous potential energy, or the energy that an object possesses by virtue of its position. Hydroelectric power is created by allowing controlled streams of this water to flow downward, gathering kinetic energy that is then transferred to powering turbines, which in turn generate electric power.

WHERE TO LEARN MORE

Aerodynamics for Students (Web site). <http://www.ae.su.oz.au/aero/contents.html> (April 8, 2001).

Beiser, Arthur. *Physics,* 5th ed. Reading, MA: Addison-Wesley, 1991.

Chahrour, Janet. *Flash! Bang! Pop! Fizz!: Exciting Science for Curious Minds.* Illustrated by Ann Humphrey Williams. Hauppauge, N.Y.: Barron's, 2000.

"Educational Fluid Mechanics Sites." *Virginia Institute of Technology* (Web site). <http://www.eng.vt.edu/fluids/links/edulinks.htm> (April 8, 2001).

Fleisher, Paul. *Liquids and Gases: Principles of Fluid Mechanics.* Minneapolis, MN: Lerner Publications, 2002.

Institute of Fluid Mechanics (Web site). <http://www.ts.go.dlr.de> (April 8, 2001).

K8AIT Principles of Aeronautics Advanced Text (web site). <http://wings.ucdavis.edu/Book/advanced.html> (February 19, 2001).

Macaulay, David. *The New Way Things Work.* Boston: Houghton Mifflin, 1998.

Sobey, Edwin J. C. *Wacky Water Fun with Science: Science You Can Float, Sink, Squirt, and Sail.* Illustrated by Bill Burg. New York: McGraw-Hill, 2000.

Wood, Robert W. *Mechanics Fundamentals.* Illustrated by Bill Wright. Philadelphia: Chelsea House, 1997.

AERODYNAMICS

CONCEPT

Though the term "aerodynamics" is most commonly associated with airplanes and the overall science of flight, in fact, its application is much broader. Simply put, aerodynamics is the study of airflow and its principles, and applied aerodynamics is the science of improving manmade objects such as airplanes and automobiles in light of those principles. Aside from the obvious application to these heavy forms of transportation, aerodynamic concepts are also reflected in the simplest of manmade flying objects—and in the natural model for all studies of flight, a bird's wings.

HOW IT WORKS

All physical objects on Earth are subject to gravity, but gravity is not the only force that tends to keep them pressed to the ground. The air itself, though it is invisible, operates in such a way as to prevent lift, much as a stone dropped into the water will eventually fall to the bottom. In fact, air behaves much like water, though the downward force is not as great due to the fact that air's pressure is much less than that of water. Yet both are media through which bodies travel, and air and water have much more in common with one another than either does with a vacuum.

Liquids such as water and gasses such as air are both subject to the principles of fluid dynamics, a set of laws that govern the motion of liquids and vapors when they come in contact with solid surfaces. In fact, there are few significant differences—for the purposes of the present discussion—between water and air with regard to their behavior in contact with solid surfaces.

When a person gets into a bathtub, the water level rises uniformly in response to the fact that a solid object is taking up space. Similarly, air currents blow over the wings of a flying aircraft in such a way that they meet again more or less simultaneously at the trailing edge of the wing. In both cases, the medium adjusts for the intrusion of a solid object. Hence within the parameters of fluid dynamics, scientists typically use the term "fluid" uniformly, even when describing the movement of air.

The study of fluid dynamics in general, and of air flow in particular, brings with it an entire vocabulary. One of the first concepts of importance is viscosity, the internal friction in a fluid that makes it resistant to flow and resistant to objects flowing through it. As one might suspect, viscosity is a far greater factor with water than with air, the viscosity of which is less than two percent that of water. Nonetheless, near a solid surface—for example, the wing of an airplane—viscosity becomes a factor because air tends to stick to that surface.

Also significant are the related aspects of density and compressibility. At speeds below 220 MPH (354 km/h), the compressibility of air is not a significant factor in aerodynamic design. However, as air flow approaches the speed of sound—660 MPH (1,622 km/h)—compressibility becomes a significant factor. Likewise temperature increases greatly when airflow is supersonic, or faster than the speed of sound.

All objects in the air are subject to two types of airflow, laminar and turbulent. Laminar flow is smooth and regular, always moving at the same speed and in the same direction. This type of airflow is also known as streamlined flow, and under these conditions every particle of fluid that passes a particular point follows a path identical to all particles that passed that point earlier. This may be illustrated by imagining a stream flowing around a twig.

By contrast, in turbulent flow the air is subject to continual changes in speed and direction—as for instance when a stream flows over shoals of rocks. Whereas the mathematical model of laminar airflow is rather straightforward, conditions are much more complex in turbulent flow, which typically occurs in the presence either of obstacles or of high speeds.

Absent the presence of viscosity, and thus in conditions of perfect laminar flow, an object behaves according to Bernoulli's principle, sometimes known as Bernoulli's equation. Named after the Swiss mathematician and physicist Daniel Bernoulli (1700-1782), this proposition goes to the heart of that which makes an airplane fly.

While conducting experiments concerning the conservation of energy in liquids, Bernoulli observed that when the diameter of a pipe is reduced, the water flows faster. This suggested to him that some force must be acting upon the water, a force that he reasoned must arise from differences in pressure. Specifically, the slower-moving fluid had a greater pressure than the portion of the fluid moving through the narrower part of the pipe. As a result, he concluded that pressure and velocity are inversely related.

Bernoulli's principle states that for all changes in movement, the sum of static and dynamic pressure in a fluid remain the same. A fluid at rest exerts static pressure, which is the same as what people commonly mean when they say "pressure," as in "water pressure." As the fluid begins to move, however, a portion of the static pressure—proportional to the speed of the fluid—is converted to what scientists call dynamic pressure, or the pressure of movement. The greater the speed, the greater the dynamic pressure and the less the static pressure. Bernoulli's findings would prove crucial to the design of aircraft in the twentieth century, as engineers learned how to use currents of faster and slower air for keeping an airplane aloft.

Very close to the surface of an object experiencing airflow, however, the presence of viscosity plays havoc with the neat proportions of the Bernoulli's principle. Here the air sticks to the object's surface, slowing the flow of nearby air and creating a "boundary layer" of slow-moving air. At the beginning of the flow—for instance, at the leading edge of an airplane's wing—this boundary layer describes a laminar flow; but the width of the layer increases as the air moves along the surface, and at some point it becomes turbulent.

A TYPICAL PAPER AIRPLANE HAS LOW ASPECT RATIO WINGS, A TERM THAT REFERS TO THE SIZE OF THE WINGSPAN COMPARED TO THE CHORD. IN SUBSONIC FLIGHT, HIGHER ASPECT RATIOS ARE USUALLY PREFERRED. *(Photograph by Bruce Burkhardt/Corbis. Reproduced by permission.)*

These and a number of other factors contribute to the coefficients of drag and lift. Simply put, drag is the force that opposes the forward motion of an object in airflow, whereas lift is a force perpendicular to the direction of the wind, which keeps the object aloft. Clearly these concepts can be readily applied to the operation of an airplane, but they also apply in the case of an automobile, as will be shown later.

REAL-LIFE APPLICATIONS

How a Bird Flies—and Why a Human Being Cannot

Birds are exquisitely designed (or adapted) for flight, and not simply because of the obvious fact

BIRDS LIKE THESE FAIRY TERNS ARE SUPREME EXAMPLES OF AERODYNAMIC PRINCIPLES, FROM THEIR LOW BODY WEIGHT AND LARGE STERNUM AND PECTORALIS MUSCLES TO THEIR LIGHTWEIGHT FEATHERS. *(Corbis. Reproduced by permission.)*

that they have wings. Thanks to light, hollow bones, their body weight is relatively low, giving them the advantage in overcoming gravity and remaining aloft. Furthermore, a bird's sternum or breast bone, as well as its pectoralis muscles (those around the chest) are enormous in proportion to its body size, thus helping it to achieve the thrust necessary for flight. And finally, the bird's lightweight feathers help to provide optimal lift and minimal drag.

A bird's wing is curved along the top, a crucial aspect of its construction. As air passes over the leading edge of the wing, it divides, and because of the curve, the air on top must travel a greater distance before meeting the air that flowed across the bottom. The tendency of airflow, as noted earlier, is to correct for the presence of solid objects. Therefore, in the absence of outside factors such as viscosity, the air on top "tries" to travel over the wing in the same amount of time that it takes the air below to travel under the wing. As shown by Bernoulli, the fast-moving air above the wing exerts less pressure than the slow-moving air below it; hence there is a difference in pressure between the air below and the air above, and this keeps the wing aloft.

When a bird beats its wings, its downstrokes propel it, and as it rises above the ground, the force of aerodynamic lift helps push its wings upward in preparation for the next downstroke. However, to reduce aerodynamic drag during the upstroke, the bird folds its wings, thus decreasing its wingspan. Another trick that birds execute instinctively is the moving of their wings forward and backward in order to provide balance. They also "know" how to flap their wings in a direction almost parallel to the ground when they need to fly slowly or hover.

Witnessing the astonishing aerodynamic feats of birds, humans sought the elusive goal of flight from the earliest of times. This was symbolized by the Greek myth of Icarus and Daedalus, who escaped from a prison in Crete by constructing a set of bird-like wings and flying away. In the world of physical reality, however, the goal would turn out to be unattainable as long as humans attempted to achieve flight by imitating birds.

As noted earlier, a bird's physiology is quite different from that of a human being. There is simply no way that a human can fly by flapping his arms—nor will there ever be a man strong enough to do so, no matter how apparently well-designed his mechanical wings are. Indeed, to be capable of flying like a bird, a man would have to have a chest so enormous in proportion to his body that he would be hideous in appearance.

Not realizing this, humans for centuries attempted to fly like birds—with disastrous results. An English monk named Eilmer (b. 980) attempted to fly off the tower of Malmesbury Abbey with a set of wings attached to his arms and feet. Apparently Eilmer panicked after gliding some 600 ft (about 200 m) and suddenly plummeted to earth, breaking both of his legs. At least he lived; more tragic was the case of Abul Qasim Ibn Firnas (d. 873), an inventor from Cordoba in Arab Spain who devised and demonstrated a glider. Much of Cordoba's population came out to see him demonstrate his flying machine, but after covering just a short distance, the craft fell to earth. Severely wounded, Ibn Firnas died shortly afterward.

The first real progress in the development of flying machines came when designers stopped trying to imitate birds and instead used the principle of buoyancy. Hence in 1783, the French brothers Jacques-Etienne and Joseph-Michel Montgolfier constructed the first practical balloon.

Balloons and their twentieth-century descendant, the dirigible, had a number of obvious drawbacks, however. Without a motor, a balloon could not be guided, and even with a motor, dirigibles proved highly dangerous. At that stage, most dirigibles used hydrogen, a gas that is cheap and plentiful, but extremely flammable. After the *Hindenburg* exploded in 1937, the age of passenger travel aboard airships was over.

However, the German military continued to use dirigibles for observation purposes, as did the United States forces in World War II. Today airships, the most famous example being the Goodyear Blimp, are used not only for observation but for advertising. Scientists working in rain forests, for instance, use dirigibles to glide above the forest canopy; as for the Goodyear Blimp, it provides television networks with "eye in the sky" views of large sporting events.

The first man to make a serious attempt at creating a heavier-than-air flying machine (as opposed to a balloon, which uses gases that are lighter than air) was Sir George Cayley (1773-1857), who in 1853 constructed a glider. It is interesting to note that in creating this, the forerunner of the modern airplane, Cayley went back to an old model: the bird. After studying the physics of birds' flight for many years, he equipped his glider with an extremely wide wingspan, used the lightest possible materials in its construction, and designed it with exceptionally smooth surfaces to reduce drag.

The only thing that in principle differentiated Cayley's craft from a modern airplane was its lack of an engine. In those days, the only possible source of power was a steam engine, which would have added far too much weight to his aircraft. However, the development of the internal-combustion engine in the nineteenth century overcame that obstacle, and in 1903 Orville and Wilbur Wright achieved the dream of flight that had intrigued and eluded human beings for centuries.

AIRPLANES: GETTING ALOFT, STAYING ALOFT, AND REMAINING STABLE

Once engineers and pilots took to the air, they encountered a number of factors that affect flight. In getting aloft and staying aloft, an aircraft is subject to weight, lift, drag, and thrust.

As noted earlier, the design of an airplane wing takes advantage of Bernoulli's principle to give it lift. Seen from the end, the wing has the shape of a long teardrop lying on its side, with the large end forward, in the direction of airflow, and the narrow tip pointing toward the rear. (Unlike a teardrop, however, an airplane's wing is asymmetrical, and the bottom side is flat.) This cross-section is known as an airfoil, and the greater curvature of its upper surface in comparison to the lower side is referred to as the airplane's camber. The front end of the airfoil is also curved, and the chord line is an imaginary straight line connecting the spot where the air hits the front—known as the stagnation point—to the rear, or trailing edge, of the wing.

Again in accordance with Bernoulli's principle, the shape of the airflow facilitates the spread of laminar flow around it. The slower-moving currents beneath the airfoil exert greater pressure than the faster currents above it, giving lift to the aircraft.

Another parameter influencing the lift coefficient (that is, the degree to which the aircraft experiences lift) is the size of the wing: the longer the wing, the greater the total force exerted beneath it, and the greater the ratio of this pressure to that of the air above. The size of a modern aircraft's wing is actually somewhat variable, due to the presence of flaps at the trailing edge.

With regard to the flaps, however, it should be noted that they have different properties at different stages of flight: in takeoff, they provide lift, but in stable flight they increase drag, and for that reason the pilot retracts them. In preparing for landing, as the aircraft slows and descends, the extended flaps then provide stability and assist in the decrease of speed.

Speed, too, encourages lift: the faster the craft, the faster the air moves over the wing. The pilot affects this by increasing or decreasing the power of the engine, thus regulating the speed with which the plane's propellers turn. Another highly significant component of lift is the airfoil's angle of attack—the orientation of the airfoil with regard to the air flow, or the angle that the chord line forms with the direction of the air stream.

Up to a point, increasing the angle of attack provides the aircraft with extra lift because it moves the stagnation point from the leading edge down along the lower surface; this increases the low-pressure area of the upper surface. However, if the pilot increases the angle of attack too much, this affects the boundary layer of slow-moving air, causing the aircraft to go into a stall.

Together the engine provides the propellers with power, and this gives the aircraft thrust, or propulsive force. In fact, the propeller blades constitute miniature wings, pivoted at the center and powered by the engine to provide rotational motion. As with the wings of the aircraft, the blades have a convex forward surface and a narrow trailing edge. Also like the aircraft wings, their angle of attack (or pitch) is adjusted at different points for differing effects. In stable flight, the pilot increases the angle of attack for the propeller blades sharply as against airflow, whereas at takeoff and landing the pitch is dramatically reduced. During landing, in fact, the pilot actually reverses the direction of the propeller blades, turning them into a brake on the aircraft's forward motion—and producing that lurching sensation that a passenger experiences as the aircraft slows after touching down.

By this point there have been several examples regarding the use of the same technique alternately to provide lift or—when slowing or preparing to land—drag. This apparent inconsistency results from the fact that the characteristics of air flow change drastically from situation to situation, and in fact, air never behaves as perfectly as it does in a textbook illustration of Bernoulli's principle.

Not only is the aircraft subject to air viscosity—the air's own friction with itself—it also experiences friction drag, which results from the fact that no solid can move through a fluid without experiencing a retarding force. An even greater drag factor, accounting for one-third of that which an aircraft experiences, is induced drag. The latter results because air does not flow in perfect laminar streams over the airfoil; rather, it forms turbulent eddies and currents that act against the forward movement of the plane.

In the air, an aircraft experiences forces that tend to destabilize flight in each of three dimensions. Pitch is the tendency to rotate forward or backward; yaw, the tendency to rotate on a horizontal plane; and roll, the tendency to rotate vertically on the axis of its fuselage. Obviously, each of these is a terrifying prospect, but fortunately, pilots have a solution for each. To prevent pitching, they adjust the angle of attack of the horizontal tail at the rear of the craft. The vertical rear tail plays a part in preventing yawing, and to prevent rolling, the pilot raises the tips of the main wings so that the craft assumes a V-shape when seen from the front or back.

The above factors of lift, drag, thrust, and weight, as well as the three types of possible destabilization, affect all forms of heavier-than-air flying machines. But since the 1944 advent of jet engines, which travel much faster than piston-driven engines, planes have flown faster and faster, and today some craft such as the *Concorde* are capable of supersonic flight. In these situations, air compressibility becomes a significant issue.

Sound is transmitted by the successive compression and expansion of air. But when a plane is traveling at above Mach 1.2—the Mach number indicates the speed of an aircraft in relation to the speed of sound—there is a significant discrepancy between the speed at which sound is traveling away from the craft, and the speed at which the craft is moving away from the sound. Eventually the compressed sound waves build up, resulting in a shock wave.

Down on the ground, the shock wave manifests as a "sonic boom"; meanwhile, for the aircraft, it can cause sudden changes in pressure, density, and temperature, as well as an increase in drag and a loss of stability. To counteract this effect, designers of supersonic and hypersonic (Mach 5 and above) aircraft are altering wing design, using a much narrower airfoil and swept-back wings.

One of the pioneers in this area is Richard Whitcomb of the National Aeronautics and Space Administration (NASA). Whitcomb has designed a supercritical airfoil for a proposed hypersonic plane, which would ascend into outer space in the course of a two-hour flight—all the time needed for it to travel from Washington, D.C., to Tokyo, Japan. Before the craft can become operational, however, researchers will have to figure out ways to control temperatures and keep the plane from bursting into flame as it reenters the atmosphere.

Much of the research for improving the aerodynamic qualities of such aircraft takes place in wind tunnels. First developed in 1871, these use powerful fans to create strong air currents, and over the years the top speed in wind tunnels has been increased to accommodate testing on supersonic and hypersonic aircraft. Researchers today use helium to create wind blasts at speeds up to Mach 50.

THROWN AND FLOWN: THE AERODYNAMICS OF SMALL OBJECTS

Long before engineers began to dream of sending planes into space for transoceanic flight—about 14,000 years ago, in fact—many of the features that make an airplane fly were already present in the boomerang. It might seem backward to move from a hypersonic jet to a boomerang, but in fact, it is easier to appreciate the aerodynamics of small objects, including the kite and even the paper airplane, once one comprehends the larger picture.

There is a certain delicious irony in the fact that the first manmade object to take flight was constructed by people who never advanced beyond the Stone Age until the nineteenth century, when the Europeans arrived in Australia. As the ethnobotanist Jared Diamond showed in his groundbreaking work *Guns, Germs, and Steel: The Fates of Human Societies* (1997), this was not because the Aborigines of Australia were less intelligent than Europeans. In fact, as Diamond showed, an individual would actually have to be smarter to figure out how to survive on the limited range of plants and animals available in Australia prior to the introduction of Eurasian flora and fauna. Hence the wonder of the boomerang, one of the most ingenious inventions ever fashioned by humans in a "primitive" state.

Thousands of years before Bernoulli, the boomerang's designers created an airfoil consistent with Bernoulli's principle. The air below exerts more pressure than the air above, and this, combined with the factors of gyroscopic stability and gyroscopic precession, gives the boomerang flight.

Gyroscopic stability can be illustrated by spinning a top: the action of spinning itself keeps the top stable. Gyroscopic precession is a much more complex process: simply put, the leading wing of the boomerang—the forward or upward edge as it spins through the air—creates more lift than the other wing. At this point it should be noted that, contrary to the popular image, a boomerang travels on a plane perpendicular to that of the ground, not parallel. Hence any thrower who knows what he or she is doing tosses the boomerang not with a side-arm throw, but overhand.

And of course a boomerang does not just sail through the air; a skilled thrower can make it come back as if by magic. This is because the force of the increased lift that it experiences in flight, combined with gyroscopic precession, turns it around. As noted earlier, in different situations the same force that creates lift can create drag, and as the boomerang spins downward the increasing drag slows it. Certainly it takes great skill for a thrower to make a boomerang come back, and for this reason, participants in boomerang competitions often attach devices such as flaps to increase drag on the return cycle.

Another very early example of an aerodynamically sophisticated humanmade device—though it is quite recent compared to the boomerang—is the kite, which first appeared in China in about 1000 B.C. The kite's design borrows from avian anatomy, particularly the bird's light, hollow bones. Hence a kite, in its simplest form, consists of two crossed strips of very light wood such as balsa, with a lightweight fabric stretched over them.

Kites can come in a variety of shapes, though for many years the well-known diamond shape has been the most popular, in part because its aerodynamic qualities make it easiest for the novice kite-flyer to handle. Like birds and boomerangs, kites can "fly" because of the physical laws embodied in Bernoulli's principle: at the best possible angle of attack, the kite experiences a maximal ratio of pressure from the slower-moving air below as against the faster-moving air above.

For centuries, when the kite represented the only way to put a humanmade object many hundreds of feet into the air, scientists and engineers used them for a variety of experiments. Of course, the most famous example of this was Benjamin Franklin's 1752 experiment with electricity. More significant to the future of aerodynamics were investigations made half a century later by Cayley, who recognized that the kite, rather than the balloon, was an appropriate model for the type of heavier-than-air flight he intended.

In later years, engineers built larger kites capable of lifting men into the air, but the advent of the airplane rendered kites obsolete for this purpose. However, in the 1950s an American engineer named Francis Rogallo invented the flexible kite, which in turn spawned the delta wing kite used by hang gliders. During the 1960s, Domina Jolbert created the parafoil, an even more efficient device, which took nonmechanized human flight perhaps as far as it can go.

Akin to the kite, glider, and hang glider is that creation of childhood fancy, the paper airplane. In its most basic form—and paper airplane enthusiasts are capable of fairly complex designs—a paper airplane is little more than a set of wings. There are a number or reasons for this, not least the fact that in most cases, a person flying a paper airplane is not as concerned about pitch, yaw, and roll as a pilot flying with several hundred passengers on board would be.

However, when fashioning a paper airplane it is possible to add a number of design features, for instance by folding flaps upward at the tail. These become the equivalent of the elevator, a control surface along the horizontal edge of a real aircraft's tail, which the pilot rotates upward to provide stability. But as noted by Ken Blackburn, author of several books on paper airplanes, it is not necessarily the case that an airplane must have a tail; indeed, some of the most sophisticated craft in the sky today—including the fearsome B-2 "Stealth" bomber—do not have tails.

A typical paper airplane has low aspect ratio wings, a term that refers to the size of the wingspan compared to the chord line. In subsonic flight, higher aspect ratios are usually preferred, and this is certainly the case with most "real" gliders; hence their wings are longer, and their chord lines shorter. But there are several reasons why this is not the case with a paper airplane.

First of all, as Blackburn noted wryly on his Web site, "Paper is a lousy building material. There is a reason why real airplanes are not made of paper." He stated the other factors governing paper airplanes' low aspect ratio in similarly whimsical terms. First, "Low aspect ratio wings are easier to fold...."; second, "Paper airplane gliding performance is not usually very important...."; and third, "Low-aspect ratio wings look faster, especially if they are swept back."

The reason why low-aspect ratio wings look faster, Blackburn suggested, is that people see them on jet fighters and the *Concorde,* and assume that a relatively narrow wing span with a long chord line yields the fastest speeds. And indeed they do—but only at supersonic speeds. Below the speed of sound, high-aspect ratio wings are best for preventing drag. Furthermore, as Blackburn went on to note, low-aspect ratio wings help the paper airplane to withstand the relatively high launch speeds necessary to send them into longer glides.

In fact, a paper airplane is not subject to anything like the sort of design constraints affecting a real craft. All real planes look somewhat similar, because the established combinations, ratios, and dimensions of wings, tails, and fuselage work best. Certainly there is a difference in basic appearance between subsonic and supersonic aircraft—but again, all supersonic jets have more or less the same low-aspect, swept wing. "With paper airplanes," Blackburn wrote, "it's easy to make airplanes that don't look like real airplanes" since "The mission of a paper airplane is [simply] to provide a good time for the pilot."

AERODYNAMICS ON THE GROUND

The preceding discussions of aerodynamics in action have concerned the behavior of objects off the ground. But aerodynamics is also a factor in wheeled transport on Earth's surface, whether by bicycle, automobile, or some other variation.

On a bicycle, the rider accounts for 65-80% of the drag, and therefore his or her position with regard to airflow is highly important. Thus, from as early as the 1890s, designers of racing bikes have favored drop handlebars, as well as a seat and frame that allow a crouched position. Since the 1980s, bicycle designers have worked to eliminate all possible extra lines and barriers to airflow, including the crossbar and chainstays.

A typical bicycle's wheel contains 32 or 36 cylindrical spokes, and these can affect aerodynamics adversely. As the wheel rotates, the airflow behind the spoke separates, creating turbulence and hence drag. For this reason, some of the most advanced bicycles today use either aerodynamic rims, which reduce the length of the spokes, three-spoke aerodynamic wheels, or even solid wheels.

The rider's gear can also serve to impede or enhance his velocity, and thus modern racing helmets have a streamlined shape—rather like that of an airfoil. The best riders, such as those who compete in the Olympics or the Tour de France, have bikes custom-designed to fit their own body shape.

One interesting aspect of aerodynamics where it concerns bicycle racing is the phenomenon of "drafting." Riders at the front of a pack, like riders pedaling alone, consume 30-40% more energy than do riders in the middle of a pack. The latter are benefiting from the efforts of bicyclists in front of them, who put up most of the wind resistance. The same is true for bicyclists who ride behind automobiles or motorcycles.

The use of machine-powered pace vehicles to help in achieving extraordinary speeds is far from new. Drafting off of a railroad car with specially designed aerodynamic shields, a rider in 1896 was able to exceed 60 MPH (96 km/h), a then unheard-of speed. Today the record is just under 167 MPH (267 km/h). Clearly one must be a highly skilled, powerful rider to approach anything like this speed; but design factors also come into play, and not just in the case of the pace vehicle. Just as supersonic jets are quite different from ordinary planes, super high-speed bicycles are not like the average bike; they are designed in such a way that they must be moving faster than 60 MPH before the rider can even pedal.

A professional bicycle racer's streamlined helmet and crouched position help to improve airflow, thus increasing speed. *(Photograph by Ronnen Eshel/Corbis. Reproduced by permission.)*

With regard to automobiles, as noted earlier, aerodynamics has a strong impact on body design. For this reason, cars over the years have become steadily more streamlined and aerodynamic in appearance, a factor that designers balance with aesthetic appeal. Today's Chrysler PT Cruiser, which debuted in 2000, may share outward features with 1930s and 1940s cars, but the PT Cruiser's design is much more sound aerodynamically—not least because a modern vehicle can travel much, much faster than the cars driven by previous generations.

Nowhere does the connection between aerodynamics and automobiles become more crucial than in the sport of auto racing. For race-car drivers, drag is always a factor to be avoided and counteracted by means ranging from drafting to altering the body design to reduce the airflow under the vehicle. However, as strange as it may seem, a car—like an airplane—is also subject to lift.

It was noted earlier that in some cases lift can be undesirable in an airplane (for instance, when trying to land), but it is virtually always undesirable in an automobile. The greater the speed, the greater the lift force, which increases

KEY TERMS

AERODYNAMICS: The study of air flow and its principles. Applied aerodynamics is the science of improving man-made objects in light of those principles.

AIRFOIL: The design of an airplane's wing when seen from the end, a shape intended to maximize the aircraft's response to airflow.

ANGLE OF ATTACK: The orientation of the airfoil with regard to the airflow, or the angle that the chord line forms with the direction of the air stream.

BERNOULLI'S PRINCIPLE: A proposition, credited to Swiss mathematician and physicist Daniel Bernoulli (1700-1782), which maintains that slower-moving fluid exerts greater pressure than faster-moving fluid.

CAMBER: The enhanced curvature on the upper surface of an airfoil.

CHORD LINE: The distance, along an imaginary straight line, from the stagnation point of an airfoil to the rear, or trailing edge.

DRAG: The force that opposes the forward motion of an object in airflow.

LAMINAR: A term describing a streamlined flow, in which all particles move at the same speed and in the same direction. Its opposite is turbulent flow.

LIFT: An aerodynamic force perpendicular to the direction of the wind. For an aircraft, lift is the force that raises it off the ground and keeps it aloft.

PITCH: The tendency of an aircraft in flight to rotate forward or backward; see also yaw and roll.

ROLL: The tendency of an aircraft in flight to rotate vertically on the axis of its fuselage; see also pitch and yaw.

STAGNATION POINT: The spot where airflow hits the leading edge of an airfoil.

SUPERSONIC: Faster than Mach 1, or the speed of sound—660 MPH (1,622 km/h). Speeds above Mach 5 are referred to as hypersonic.

TURBULENT: A term describing a highly irregular form of flow, in which a fluid is subject to continual changes in speed and direction. Its opposite is laminar flow.

VISCOSITY: The internal friction in a fluid that makes it resistant to flow.

YAW: The tendency of an aircraft in flight to rotate on a horizontal plane; see also Pitch and Roll.

the threat of instability. For this reason, builders of race cars design their vehicles for negative lift: hence a typical family car has a lift coefficient of about 0.03, whereas a race car is likely to have a coefficient of -3.00.

Among the design features most often used to reduce drag while achieving negative lift is a rear-deck spoiler. The latter has an airfoil shape, but its purpose is different: to raise the rear stagnation point and direct air flow so that it does not wrap around the vehicle's rear end. Instead, the spoiler creates a downward force to stabilize the rear, and it may help to decrease drag by reducing the separation of airflow (and hence the creation of turbulence) at the rear window.

Similar in concept to a spoiler, though somewhat different in purpose, is the aerodynamically curved shield that sits atop the cab of most modern eighteen-wheel transport trucks. The purpose of the shield becomes apparent when the truck is moving at high speeds: wind resistance becomes strong, and if the wind were to hit the truck's trailer head-on, it would be as though the air were pounding a brick wall. Instead, the shield scoops air upward, toward the rear of the truck. At the rear may be another panel, patented by two young engineers in 1994, that creates a drag-reducing vortex between panel and truck.

WHERE TO LEARN MORE

Cockpit Physics (Department of Physics, United States Air Force Academy web site.). <http://www.usafa.af.mil/dfp/cockpit-phys/> (February 19, 2001).

K8AIT Principles of Aeronautics Advanced Text. (web site). <http://wings.ucdavis.edu/Book/advanced.html> (February 19, 2001).

Macaulay, David. *The New Way Things Work.* Boston: Houghton Mifflin, 1998.

Blackburn, Ken. *Paper Airplane Aerodynamics.* (web site). <http://www.geocities.com/CapeCanaveral/1817/paero.html> (February 19, 2001).

Schrier, Eric and William F. Allman. *Newton at the Bat: The Science in Sports.* New York: Charles Scribner's Sons, 1984.

Smith, H. C. *The Illustrated Guide to Aerodynamics.* Blue Ridge Summit, PA: Tab Books, 1992.

Stever, H. Guyford, James J. Haggerty, and the Editors of Time-Life Books. *Flight.* New York: Time-Life Books, 1965.

Suplee, Curt. *Everyday Science Explained.* Washington, D.C.: National Geographic Society, 1996.

BERNOULLI'S PRINCIPLE

CONCEPT

Bernoulli's principle, sometimes known as Bernoulli's equation, holds that for fluids in an ideal state, pressure and density are inversely related: in other words, a slow-moving fluid exerts more pressure than a fast-moving fluid. Since "fluid" in this context applies equally to liquids and gases, the principle has as many applications with regard to airflow as to the flow of liquids. One of the most dramatic everyday examples of Bernoulli's principle can be found in the airplane, which stays aloft due to pressure differences on the surface of its wing; but the truth of the principle is also illustrated in something as mundane as a shower curtain that billows inward.

HOW IT WORKS

The Swiss mathematician and physicist Daniel Bernoulli (1700-1782) discovered the principle that bears his name while conducting experiments concerning an even more fundamental concept: the conservation of energy. This is a law of physics that holds that a system isolated from all outside factors maintains the same total amount of energy, though energy transformations from one form to another take place.

For instance, if you were standing at the top of a building holding a baseball over the side, the ball would have a certain quantity of potential energy—the energy that an object possesses by virtue of its position. Once the ball is dropped, it immediately begins losing potential energy and gaining kinetic energy—the energy that an object possesses by virtue of its motion. Since the total energy must remain constant, potential and kinetic energy have an inverse relationship: as the value of one variable decreases, that of the other increases in exact proportion.

The ball cannot keep falling forever, losing potential energy and gaining kinetic energy. In fact, it can never gain an amount of kinetic energy greater than the potential energy it possessed in the first place. At the moment before the ball hits the ground, its kinetic energy is equal to the potential energy it possessed at the top of the building. Correspondingly, its potential energy is zero—the same amount of kinetic energy it possessed before it was dropped.

Then, as the ball hits the ground, the energy is dispersed. Most of it goes into the ground, and depending on the rigidity of the ball and the ground, this energy may cause the ball to bounce. Some of the energy may appear in the form of sound, produced as the ball hits bottom, and some will manifest as heat. The total energy, however, will not be lost: it will simply have changed form.

Bernoulli was one of the first scientists to propose what is known as the kinetic theory of gases: that gas, like all matter, is composed of tiny molecules in constant motion. In the 1730s, he conducted experiments in the conservation of energy using liquids, observing how water flows through pipes of varying diameter. In a segment of pipe with a relatively large diameter, he observed, water flowed slowly, but as it entered a segment of smaller diameter, its speed increased.

It was clear that some force had to be acting on the water to increase its speed. Earlier, Robert Boyle (1627-1691) had demonstrated that pressure and volume have an inverse relationship, and Bernoulli seems to have applied Boyle's findings to the present situation. Clearly the volume of water flowing through the narrower pipe at any given moment was less than that flowing through the wider one. This suggested, according to Boyle's law, that the pressure in the wider pipe must be greater.

As fluid moves from a wider pipe to a narrower one, the volume of that fluid that moves a given distance in a given time period does not change. But since the width of the narrower pipe is smaller, the fluid must move faster in order to achieve that result. One way to illustrate this is to observe the behavior of a river: in a wide, unconstricted region, it flows slowly, but if its flow is narrowed by canyon walls (for instance), then it speeds up dramatically.

The above is a result of the fact that water is a fluid, and having the characteristics of a fluid, it adjusts its shape to fit that of its container or other solid objects it encounters on its path. Since the volume passing through a given length of pipe during a given period of time will be the same, there must be a decrease in pressure. Hence Bernoulli's conclusion: the slower the rate of flow, the higher the pressure, and the faster the rate of flow, the lower the pressure.

Bernoulli published the results of his work in *Hydrodynamica* (1738), but did not present his ideas or their implications clearly. Later, his friend the German mathematician Leonhard Euler (1707-1783) generalized his findings in the statement known today as Bernoulli's principle.

The Venturi Tube

Also significant was the work of the Italian physicist Giovanni Venturi (1746-1822), who is credited with developing the Venturi tube, an instrument for measuring the drop in pressure that takes place as the velocity of a fluid increases. It consists of a glass tube with an inward-sloping area in the middle, and manometers, devices for measuring pressure, at three places: the entrance, the point of constriction, and the exit. The Venturi meter provided a consistent means of demonstrating Bernoulli's principle.

Like many propositions in physics, Bernoulli's principle describes an ideal situation in the absence of other forces. One such force is viscosity, the internal friction in a fluid that makes it resistant to flow. In 1904, the German physicist Ludwig Prandtl (1875-1953) was conducting experiments in liquid flow, the first effort in well over a century to advance the findings of Bernoulli and others. Observing the flow of liquid in a tube, Prandtl found that a tiny portion of the liquid adheres to the surface of the tube in the form of a thin film, and does not continue to move. This he called the viscous boundary layer.

Like Bernoulli's principle itself, Prandtl's findings would play a significant part in aerodynamics, or the study of airflow and its principles. They are also significant in hydrodynamics, or the study of water flow and its principles, a discipline Bernoulli founded.

Laminar vs. Turbulent Flow

Air and water are both examples of fluids, substances which—whether gas or liquid—conform to the shape of their container. The flow patterns of all fluids may be described in terms either of laminar flow, or of its opposite, turbulent flow.

Laminar flow is smooth and regular, always moving at the same speed and in the same direction. Also known as streamlined flow, it is characterized by a situation in which every particle of fluid that passes a particular point follows a path identical to all particles that passed that point earlier. A good illustration of laminar flow is what occurs when a stream flows around a twig.

By contrast, in turbulent flow, the fluid is subject to continual changes in speed and direction—as, for instance, when a stream flows over shoals of rocks. Whereas the mathematical model of laminar flow is rather straightforward, conditions are much more complex in turbulent flow, which typically occurs in the presence of obstacles or high speeds.

Turbulent flow makes it more difficult for two streams of air, separated after hitting a barrier, to rejoin on the other side of the barrier; yet that is their natural tendency. In fact, if a single air current hits an airfoil—the design of an airplane's wing when seen from the end, a streamlined shape intended to maximize the aircraft's response to airflow—the air that flows over the top will "try" to reach the back end of the airfoil at the same time as the air that flows over the

A KITE'S DESIGN, PARTICULARLY ITS USE OF LIGHTWEIGHT FABRIC STRETCHED OVER TWO CROSSED STRIPS OF VERY LIGHT WOOD, MAKES IT WELL-SUITED FOR FLIGHT, BUT WHAT KEEPS IT IN THE AIR IS A DIFFERENCE IN AIR PRESSURE. AT THE BEST POSSIBLE ANGLE OF ATTACK, THE KITE EXPERIENCES AN IDEAL RATIO OF PRESSURE FROM THE SLOWER-MOVING AIR BELOW VERSUS THE FASTER-MOVING AIR ABOVE, AND THIS GIVES IT LIFT. *(Roger Ressmeyer/Corbis. Reproduced by permission.)*

bottom. In order to do so, it will need to speed up—and this, as will be shown below, is the basis for what makes an airplane fly.

When viscosity is absent, conditions of perfect laminar flow exist: an object behaves in complete alignment with Bernoulli's principle. Of

The boomerang, developed by Australia's Aboriginal people, flies through the air on a plane perpendicular to the ground, rather than parallel. As it flies, the boomerang becomes both a gyroscope and an airfoil, and this dual role gives it aerodynamic lift. *(Bettmann/Corbis. Reproduced by permission.)*

course, though ideal conditions seldom occur in the real world, Bernoulli's principle provides a guide for the behavior of planes in flight, as well as a host of everyday things.

REAL-LIFE APPLICATIONS

Flying Machines

For thousands of years, human beings vainly sought to fly "like a bird," not realizing that this is literally impossible, due to differences in physiognomy between birds and *homo sapiens*. No man has ever been born (or ever will be) who possesses enough strength in his chest that he could flap a set of attached wings and lift his body off the ground. Yet the bird's physical structure proved highly useful to designers of practical flying machines.

A bird's wing is curved along the top, so that when air passes over the wing and divides, the curve forces the air on top to travel a greater distance than the air on the bottom. The tendency of airflow, as noted earlier, is to correct for the presence of solid objects and to return to its original pattern as quickly as possible. Hence, when the air hits the front of the wing, the rate of flow at the top increases to compensate for the greater distance it has to travel than the air below the wing. And as shown by Bernoulli, fast-moving fluid exerts less pressure than slow-moving fluid; therefore, there is a difference in pressure between the air below and the air above, and this keeps the wing aloft.

Only in 1853 did Sir George Cayley (1773-1857) incorporate the avian airfoil to create history's first workable (though engine-less) flying machine, a glider. Much, much older than Cayley's glider, however, was the first manmade flying machine built "according to Bernoulli's principle"—only it first appeared in about 12,000 B.C., and the people who created it had had little contact with the outside world until the late eighteenth century A.D. This was the boomerang, one of the most ingenious devices ever created by a stone-age society—in this case, the Aborigines of Australia.

Contrary to the popular image, a boomerang flies through the air on a plane perpendicular to the ground, rather than parallel. Hence, any thrower who properly knows how tosses the boomerang not with a side-arm throw, but overhand. As it flies, the boomerang becomes both a gyroscope and an airfoil, and this dual role gives it aerodynamic lift.

Like the gyroscope, the boomerang imitates a top; spinning keeps it stable. It spins through the air, its leading wing (the forward or upward wing) creating more lift than the other wing. As an airfoil, the boomerang is designed so that the air below exerts more pressure than the air above, which keeps it airborne.

Another very early example of a flying machine using Bernoulli's principles is the kite, which first appeared in China in about 1000 B.C. The kite's design, particularly its use of lightweight fabric stretched over two crossed strips of very light wood, makes it well-suited for flight, but what keeps it in the air is a difference in air pressure. At the best possible angle of attack, the kite experiences an ideal ratio of pressure from the slower-moving air below versus the faster-moving air above, and this gives it lift.

Later Cayley studied the operation of the kite, and recognized that it—rather than the balloon, which at first seemed the most promising apparatus for flight—was an appropriate model for the type of heavier-than-air flying machine he intended to build. Due to the lack of a motor, however, Cayley's prototypical airplane could never be more than a glider: a steam engine, then state-of-the-art technology, would have been much too heavy.

Hence, it was only with the invention of the internal-combustion engine that the modern airplane came into being. On December 17, 1903, at Kitty Hawk, North Carolina, Orville (1871-1948) and Wilbur (1867-1912) Wright tested a craft that used a 25-horsepower engine they had developed at their bicycle shop in Ohio. By maximizing the ratio of power to weight, the engine helped them overcome the obstacles that had dogged recent attempts at flight, and by the time the day was over, they had achieved a dream that had eluded men for more than four millennia.

Within fifty years, airplanes would increasingly obtain their power from jet rather than internal-combustion engines. But the principle that gave them flight, and the principle that kept them aloft once they were airborne, reflected back to Bernoulli's findings of more than 160 years before their time. This is the concept of the airfoil.

As noted earlier, an airfoil has a streamlined design. Its shape is rather like that of an elongated, asymmetrical teardrop lying on its side, with the large end toward the direction of airflow, and the narrow tip pointing toward the rear. The greater curvature of its upper surface in comparison to the lower side is referred to as the airplane's camber. The front end of the airfoil is also curved, and the chord line is an imaginary straight line connecting the spot where the air hits the front—known as the stagnation point—to the rear, or trailing edge, of the wing.

Again, in accordance with Bernoulli's principle, the shape of the airflow facilitates the spread of laminar flow around it. The slower-moving currents beneath the airfoil exert greater pressure than the faster currents above it, giving lift to the aircraft. Of course, the aircraft has to be moving at speeds sufficient to gain momentum for its leap from the ground into the air, and here again, Bernoulli's principle plays a part.

Thrust comes from the engines, which run the propellers—whose blades in turn are designed as miniature airfoils to maximize their power by harnessing airflow. Like the aircraft wings, the blades' angle of attack—the angle at which airflow hits it. In stable flight, the pilot greatly increases the angle of attack (also called pitched), whereas at takeoff and landing, the pitch is dramatically reduced.

Drawing Fluids Upward: Atomizers and Chimneys

A number of everyday objects use Bernoulli's principle to draw fluids upward, and though in terms of their purposes, they might seem very different—for instance, a perfume atomizer vs. a chimney—they are closely related in their application of pressure differences. In fact, the idea behind an atomizer for a perfume spray bottle can also be found in certain garden-hose attachments, such as those used to provide a high-pressure car wash.

The air inside the perfume bottle is moving relatively slowly; therefore, according to Bernoulli's principle, its pressure is relatively high, and it exerts a strong downward force on the perfume itself. In an atomizer there is a narrow tube running from near the bottom of the bottle to the top. At the top of the perfume bottle, it opens inside another tube, this one perpendicular to the first tube. At one end of the horizontal tube is a simple squeeze-pump which causes air to flow quickly through it. As a result, the pressure toward the top of the bottle is reduced, and the perfume flows upward along the vertical tube, drawn from the area of higher pressure at the bottom. Once it is in the upper tube, the squeeze-pump helps to eject it from the spray nozzle.

A carburetor works on a similar principle, though in that case the lower pressure at the top draws air rather than liquid. Likewise a chimney draws air upward, and this explains why a windy day outside makes for a better fire inside. With wind blowing over the top of the chimney, the air pressure at the top is reduced, and tends to draw higher-pressure air from down below.

The upward pull of air according to the Bernoulli principle can also be illustrated by what is sometimes called the "Hoover bugle"—a name perhaps dating from the Great Depression, when anything cheap or contrived bore the appellation "Hoover" as a reflection of popular dissatisfaction with President Herbert Hoover. In any case, the Hoover bugle is simply a long corrugated tube that, when swung overhead, produces musical notes.

You can create a Hoover bugle using any sort of corrugated tube, such as vacuum-cleaner hose or swimming-pool drain hose, about 1.8 in (4 cm) in diameter and 6 ft (1.8 m) in length. To operate it, you should simply hold the tube in both hands, with extra length in the leading hand—that is, the right hand, for most people. This is the hand with which to swing the tube over your head, first slowly and then faster, observing the changes in tone that occur as you change the pace.

The vacuum hose of a Hoover tube can also be returned to a version of its original purpose in an illustration of Bernoulli's principle. If a piece of paper is torn into pieces and placed on a table, with one end of the tube just above the paper and the other end spinning in the air, the paper tends to rise. It is drawn upward as though by a vacuum cleaner—but in fact, what makes it happen is the pressure difference created by the movement of air.

In both cases, reduced pressure draws air from the slow-moving region at the bottom of the tube. In the case of the Hoover bugle, the corrugations produce oscillations of a certain frequency. Slower speeds result in slower oscillations and hence lower frequency, which produces a lower tone. At higher speeds, the opposite is true. There is little variation in tones on a Hoover bugle: increasing the velocity results in a frequency twice that of the original, but it is difficult to create enough speed to generate a third tone.

Spin, Curve, and Pull: The Counterintuitive Principle

There are several other interesting illustrations—sometimes fun and in one case potentially tragic—of Bernoulli's principle. For instance, there is the reason why a shower curtain billows inward once the shower is turned on. It would seem logical at first that the pressure created by the water would push the curtain outward, securing it to the side of the bathtub.

Instead, of course, the fast-moving air generated by the flow of water from the shower creates a center of lower pressure, and this causes the curtain to move away from the slower-moving air outside. This is just one example of the ways in which Bernoulli's principle creates results that, on first glance at least, seem counterintuitive—that is, the opposite of what common sense would dictate.

Another fascinating illustration involves placing two empty soft drink cans parallel to one another on a table, with a couple of inches or a few centimeters between them. At that point, the air on all sides has the same slow speed. If you were to blow directly between the cans, however, this would create an area of low pressure between them. As a result, the cans push together. For ships in a harbor, this can be a frightening prospect: hence, if two crafts are parallel to one another and a strong wind blows between them, there is a possibility that they may behave like the cans.

Then there is one of the most illusory uses of Bernoulli's principle, that infamous baseball pitcher's trick called the curve ball. As the ball moves through the air toward the plate, its velocity creates an air stream moving against the trajectory of the ball itself. Imagine it as two lines, one curving over the ball and one curving under, as the ball moves in the opposite direction.

In an ordinary throw, the effects of the airflow would not be particularly intriguing, but in this case, the pitcher has deliberately placed a "spin" on the ball by the manner in which he has thrown it. How pitchers actually produce spin is a complex subject unto itself, involving grip,

KEY TERMS

AERODYNAMICS: The study of airflow and its principles. Applied aerodynamics is the science of improving man-made objects in light of those principles.

AIRFOIL: The design of an airplane's wing when seen from the end, a shape intended to maximize the aircraft's response to airflow.

ANGLE OF ATTACK: The orientation of the airfoil with regard to the airflow, or the angle that the chord line forms with the direction of the air stream.

BERNOULLI'S PRINCIPLE: A proposition, credited to Swiss mathematician and physicist Daniel Bernoulli (1700-1782), which maintains that slower-moving fluid exerts greater pressure than faster-moving fluid.

CAMBER: The enhanced curvature on the upper surface of an airfoil.

CHORD LINE: The distance, along an imaginary straight line, from the stagnation point of an airfoil to the rear, or trailing edge.

CONSERVATION OF ENERGY: A law of physics which holds that within a system isolated from all other outside factors, the total amount of energy remains the same, though transformations of energy from one form to another take place.

FLUID: Any substance, whether gas or liquid, that conforms to the shape of its container.

HYDRODYNAMICS: The study of water flow and its principles.

INVERSE RELATIONSHIP: A situation involving two variables, in which one of the two increases in direct proportion to the decrease in the other.

KINETIC ENERGY: The energy that an object possesses by virtue of its motion.

LAMINAR: A term describing a streamlined flow, in which all particles move at the same speed and in the same direction. Its opposite is turbulent flow.

LIFT: An aerodynamic force perpendicular to the direction of the wind. For an aircraft, lift is the force that raises it off the ground and keeps it aloft.

MANOMETERS: Devices for measuring pressure in conjunction with a Venturi tube.

POTENTIAL ENERGY: The energy that an object possesses by virtue of its position.**STAGNATION POINT:** The spot where airflow hits the leading edge of an airfoil.

TURBULENT: A term describing a highly irregular form of flow, in which a fluid is subject to continual changes in speed and direction. Its opposite is laminar flow.

VENTURI TUBE: An instrument, consisting of a glass tube with an inward-sloping area in the middle, for measuring the drop in pressure that takes place as the velocity of a fluid increases.

VISCOSITY: The internal friction in a fluid that makes it resistant to flow.

wrist movement, and other factors, and in any case, the fact of the spin is more important than the way in which it was achieved.

If the direction of airflow is from right to left, the ball, as it moves into the airflow, is spinning clockwise. This means that the air flowing over the ball is moving in a direction opposite to the spin, whereas that flowing under it is moving in the same direction. The opposite forces produce a drag on the top of the ball, and this cuts down on the velocity at the top compared to that at the bottom of the ball, where spin and airflow are moving in the same direction.

Thus the air pressure is higher at the top of the ball, and as per Bernoulli's principle, this tends to pull the ball downward. The curve ball—of which there are numerous variations, such as the fade and the slider—creates an unpredictable situation for the batter, who sees the ball leave the pitcher's hand at one altitude, but finds to his dismay that it has dropped dramatically by the time it crosses the plate.

A final illustration of Bernoulli's often counterintuitive principle neatly sums up its effects on the behavior of objects. To perform the experiment, you need only an index card and a flat surface. The index card should be folded at the ends so that when the card is parallel to the surface, the ends are perpendicular to it. These folds should be placed about half an inch (about one centimeter) from the ends.

At this point, it would be handy to have an unsuspecting person—someone who has not studied Bernoulli's principle—on the scene, and challenge him or her to raise the card by blowing under it. Nothing could seem easier, of course: by blowing under the card, any person would naturally assume, the air will lift it. But of course this is completely wrong according to Bernoulli's principle. Blowing under the card, as illustrated, will create an area of high velocity and low pressure. This will do nothing to lift the card: in fact, it only pushes the card more firmly down on the table.

WHERE TO LEARN MORE

Beiser, Arthur. *Physics,* 5th ed. Reading, MA: Addison-Wesley, 1991.

"*Bernoulli's Principle: Explanations and Demos.*" (Web site). <http://207.10.97.102/physicszone/lesson/02forces/bernoull/bernoul l.html> (February 22, 2001).

Cockpit Physics (Department of Physics, United States Air Force Academy. Web site.). <http://www.usafa.af.mil/dfp/cockpit-phys/> (February 19, 2001).

K8AIT Principles of Aeronautics Advanced Text. (Web site). <http://wings.ucdavis.edu/Book/advanced.html> (February 19, 2001).

Schrier, Eric and William F. Allman. *Newton at the Bat: The Science in Sports.* New York: Charles Scribner's Sons, 1984.

Smith, H. C. *The Illustrated Guide to Aerodynamics.* Blue Ridge Summit, PA: Tab Books, 1992.

Stever, H. Guyford, James J. Haggerty, and the Editors of Time-Life Books. *Flight.* New York: Time-Life Books, 1965.

BUOYANCY

CONCEPT

The principle of buoyancy holds that the buoyant or lifting force of an object submerged in a fluid is equal to the weight of the fluid it has displaced. The concept is also known as Archimedes's principle, after the Greek mathematician, physicist, and inventor Archimedes (c. 287-212 B.C.), who discovered it. Applications of Archimedes's principle can be seen across a wide vertical spectrum: from objects deep beneath the oceans to those floating on its surface, and from the surface to the upper limits of the stratosphere and beyond.

HOW IT WORKS

Archimedes Discovers Buoyancy

There is a famous story that Sir Isaac Newton (1642-1727) discovered the principle of gravity when an apple fell on his head. The tale, an exaggerated version of real events, has become so much a part of popular culture that it has been parodied in television commercials. Almost equally well known is the legend of how Archimedes discovered the concept of buoyancy.

A native of Syracuse, a Greek colony in Sicily, Archimedes was related to one of that city's kings, Hiero II (308?-216 B.C.). After studying in Alexandria, Egypt, he returned to his hometown, where he spent the remainder of his life. At some point, the royal court hired (or compelled) him to set about determining the weight of the gold in the king's crown. Archimedes was in his bath pondering this challenge when suddenly it occurred to him that the buoyant force of a submerged object is equal to the weight of the fluid displaced by it.

He was so excited, the legend goes, that he jumped out of his bath and ran naked through the streets of Syracuse shouting "Eureka!" (I have found it). Archimedes had recognized a principle of enormous value—as will be shown—to shipbuilders in his time, and indeed to shipbuilders of the present.

Concerning the history of science, it was a particularly significant discovery; few useful and enduring principles of physics date to the period before Galileo Galilei (1564-1642.) Even among those few ancient physicists and inventors who contributed work of lasting value—Archimedes, Hero of Alexandria (c. 65-125 A.D.), and a few others—there was a tendency to miss the larger implications of their work. For example, Hero, who discovered steam power, considered it useful only as a toy, and as a result, this enormously significant discovery was ignored for seventeen centuries.

In the case of Archimedes and buoyancy, however, the practical implications of the discovery were more obvious. Whereas steam power must indeed have seemed like a fanciful notion to the ancients, there was nothing farfetched about oceangoing vessels. Shipbuilders had long been confronted with the problem of how to keep a vessel afloat by controlling the size of its load on the one hand, and on the other hand, its tendency to bob above the water. Here, Archimedes offered an answer.

Buoyancy and Weight

Why does an object seem to weigh less underwater than above the surface? How is it that a ship made of steel, which is obviously heavier than water, can float? How can we determine whether a balloon will ascend in the air, or a submarine will descend in the water? These and other questions are addressed by the principle of buoyancy, which can be explained in terms of properties—most notably, gravity—unknown to Archimedes.

To understand the factors at work, it is useful to begin with a thought experiment. Imagine a certain quantity of fluid submerged within a larger body of the same fluid. Note that the terms "liquid" or "water" have not been used: not only is "fluid" a much more general term, but also, in general physical terms and for the purposes of the present discussion, there is no significant difference between gases and liquids. Both conform to the shape of the container in which they are placed, and thus both are fluids.

To return to the thought experiment, what has been posited is in effect a "bag" of fluid—that is, a "bag" made out of fluid and containing fluid no different from the substance outside the "bag." This "bag" is subjected to a number of forces. First of all, there is its weight, which tends to pull it to the bottom of the container. There is also the pressure of the fluid all around it, which varies with depth: the deeper within the container, the greater the pressure.

Pressure is simply the exertion of force over a two-dimensional area. Thus it is as though the fluid is composed of a huge number of two-dimensional "sheets" of fluid, each on top of the other, like pages in a newspaper. The deeper into the larger body of fluid one goes, the greater the pressure; yet it is precisely this increased force at the bottom of the fluid that tends to push the "bag" upward, against the force of gravity.

Now consider the weight of this "bag." Weight is a force—the product of mass multiplied by acceleration—that is, the downward acceleration due to Earth's gravitational pull. For an object suspended in fluid, it is useful to substitute another term for mass. Mass is equal to volume, or the amount of three-dimensional space occupied by an object, multiplied by density. Since density is equal to mass divided by volume, this means that volume multiplied by density is the same as mass.

We have established that the weight of the fluid "bag" is Vdg, where V is volume, d is density, and g is the acceleration due to gravity. Now imagine that the "bag" has been replaced by a solid object of exactly the same size. The solid object will experience exactly the same degree of pressure as the imaginary "bag" did—and hence, it will also experience the same buoyant force pushing it up from the bottom. This means that buoyant force is equal to the weight—Vdg—of displaced fluid.

Buoyancy is always a double-edged proposition. If the buoyant force on an object is greater than the weight of that object—in other words, if the object weighs less than the amount of water it has displaced—it will float. But if the buoyant force is less than the object's weight, the object will sink. Buoyant force is not the same as net force: if the object weighs more than the water it displaces, the force of its weight cancels out and in fact "overrules" that of the buoyant force.

At the heart of the issue is density. Often, the density of an object in relation to water is referred to as its specific gravity: most metals, which are heavier than water, are said to have a high specific gravity. Conversely, petroleum-based products typically float on the surface of water, because their specific gravity is low. Note the close relationship between density and weight where buoyancy is concerned: in fact, the most buoyant objects are those with a relatively high volume and a relatively low density.

This can be shown mathematically by means of the formula noted earlier, whereby density is equal to mass divided by volume. If $Vd = V(m/V)$, an increase in density can only mean an increase in mass. Since weight is the product of mass multiplied by g (which is assumed to be a constant figure), then an increase in density means an increase in mass and hence, an increase in weight—not a good thing if one wants an object to float.

REAL-LIFE APPLICATIONS

Staying Afloat

In the early 1800s, a young Mississippi River flatboat operator submitted a patent application describing a device for "buoying vessels over shoals." The invention proposed to prevent a problem he had often witnessed on the river—boats grounded on sandbars—by equipping the boats with adjustable buoyant air chambers. The young man even whittled a model of his invention, but he was not destined for fame as an inventor; instead, Abraham Lincoln (1809-1865) was famous for much else. In fact Lincoln had a sound idea with his proposal to use buoyant force in protecting boats from running aground.

Buoyancy on the surface of water has a number of easily noticeable effects in the real world. (Having established the definition of fluid, from this point onward, the fluids discussed will be primarily those most commonly experienced: water and air.) It is due to buoyancy that fish, human swimmers, icebergs, and ships stay afloat. Fish offer an interesting application of volume change as a means of altering buoyancy: a fish has an internal swim bladder, which is filled with gas. When it needs to rise or descend, it changes the volume in its swim bladder, which then changes its density. The examples of swimmers and icebergs directly illustrate the principle of density—on the part of the water in the first instance, and on the part of the object itself in the second.

To a swimmer, the difference between swimming in fresh water and salt water shows that buoyant force depends as much on the density of the fluid as on the volume displaced. Fresh water has a density of 62.4 lb/ft^3 (9,925 N/m^3), whereas that of salt water is 64 lb/ft^3 (10,167 N/m^3). For this reason, salt water provides more buoyant force than fresh water; in Israel's Dead Sea, the saltiest body of water on Earth, bathers experience an enormous amount of buoyant force.

Water is an unusual substance in a number of regards, not least its behavior as it freezes. Close to the freezing point, water thickens up, but once it turns to ice, it becomes less dense. This is why ice cubes and icebergs float. However, their low density in comparison to the water around them means that only part of an iceberg stays atop the surface. The submerged percentage of an iceberg is the same as the ratio of the density of ice to that of water: 89%.

Ships at Sea

Because water itself is relatively dense, a high-volume, low-density object is likely to displace a quantity of water more dense—and heavier—than the object itself. By contrast, a steel ball dropped into the water will sink straight to the bottom, because it is a low-volume, high-density object that outweighs the water it displaced.

This brings back the earlier question: how can a ship made out of steel, with a density of 487 lb/ft^3 (77,363 N/m^3), float on a salt-water ocean with an average density of only about one-eighth that amount? The answer lies in the design of the ship's hull. If the ship were flat like a raft, or if all the steel in it were compressed into a ball, it would indeed sink. Instead, however, the hollow hull displaces a volume of water heavier than the ship's own weight: once again, volume has been maximized, and density minimized.

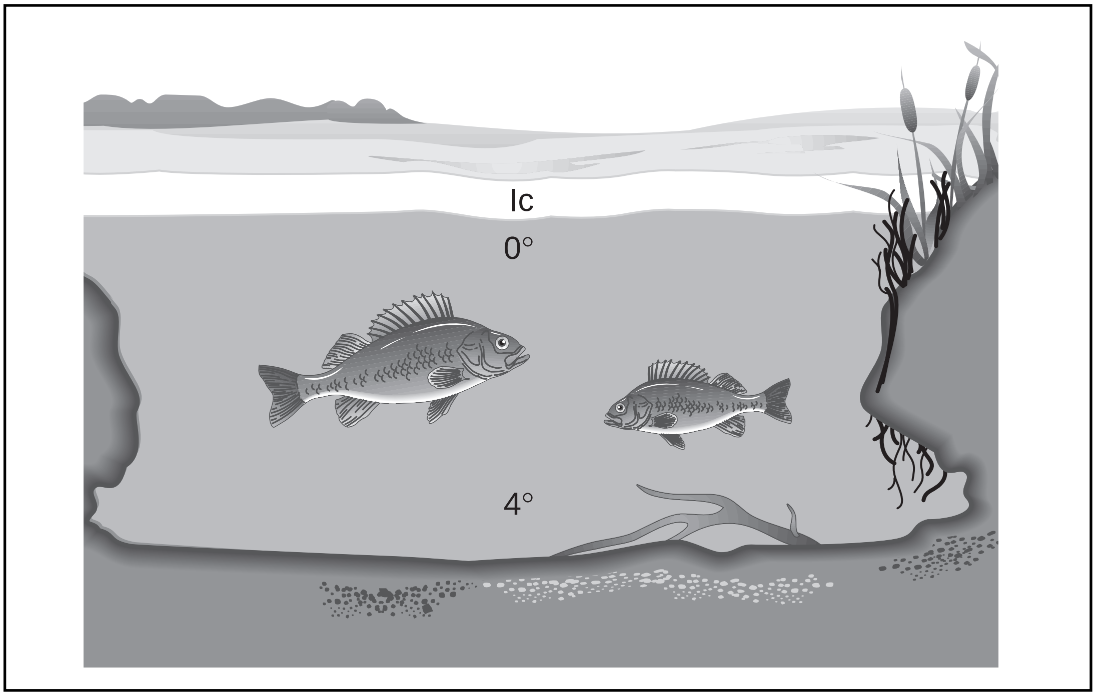
For a ship to be seaworthy, it must maintain a delicate balance between buoyancy and stability. A vessel that is too light—that is, too much volume and too little density—will bob on the top of the water. Therefore, it needs to carry a certain amount of cargo, and if not cargo, then water or some other form of ballast. Ballast is a heavy substance that increases the weight of an object experiencing buoyancy, and thereby improves its stability.

Ideally, the ship's center of gravity should be vertically aligned with its center of buoyancy. The center of gravity is the geometric center of the ship's weight—the point at which weight above is equal to weight below, weight fore is equal to weight aft, and starboard (right-side) weight is equal to weight on the port (left) side. The center of buoyancy is the geometric center of its submerged volume, and in a stable ship, it is some distance directly below center of gravity.

Displacement, or the weight of the fluid that is moved out of position when an object is immersed, gives some idea of a ship's stability. If a ship set down in the ocean causes 1,000 tons ($8.896 \cdot 10^6$ N) of water to be displaced, it is said to possess a displacement of 1,000 tons. Obviously, a high degree of displacement is desirable. The principle of displacement helps to explain how an aircraft carrier can remain afloat, even though it weighs many thousands of tons.

Down to the Depths

A submarine uses ballast as a means of descending and ascending underwater: when the submarine captain orders the crew to take the craft down, the craft is allowed to take water into its ballast tanks. If, on the other hand, the command is given to rise toward the surface, a valve will be opened to release compressed air into the tanks. The air pushes out the water, and causes the craft to ascend.



The molecular structure of water begins to expand once it cools beyond 39.4°F (4°C) and continues to expand until it becomes ice. For this reason, ice is less dense than water, floats on the surface, and retards further cooling of deeper water, which accounts for the survival of freshwater plant and animal life through the winter. For their part, fish change the volume of their internal swim bladder in order to alter their buoyancy.

A submarine is an underwater ship; its streamlined shape is designed to ease its movement. On the other hand, there are certain kinds of underwater vessels, known as submersibles, that are designed to sink—in order to observe or collect data from the ocean floor. Originally, the idea of a submersible was closely linked to that of diving itself. An early submersible was the diving bell, a device created by the noted English astronomer Edmund Halley (1656-1742.)

Though his diving bell made it possible for Halley to set up a company in which hired divers salvaged wrecks, it did not permit divers to go beyond relatively shallow depths. First of all, the diving bell received air from the surface: in Halley's time, no technology existed for taking an oxygen supply below. Nor did it provide substantial protection from the effects of increased pressure at great depths.

PERILS OF THE DEEP. The most immediate of those effects is, of course, the tendency of an object experiencing such pressure to simply implode like a tin can in a vise. Furthermore, the human body experiences several severe reactions to great depth: under water, nitrogen gas accumulates in a diver's bodily tissues, producing two different—but equally frightening—effects.

Nitrogen is an inert gas under normal conditions, yet in the high pressure of the ocean depths it turns into a powerful narcotic, causing nitrogen narcosis—often known by the poetic-sounding name "rapture of the deep." Under the influence of this deadly euphoria, divers begin to think themselves invincible, and their altered judgment can put them into potentially fatal situations.

Nitrogen narcosis can occur at depths as shallow as 60 ft (18.29 m), and it can be overcome simply by returning to the surface. However, one should not return to the surface too quickly, particularly after having gone down to a significant depth for a substantial period of time. In such an instance, on returning to the surface nitrogen gas will bubble within the body, producing decompression sickness—known colloquially as "the bends." This condition may manifest as itching and other skin problems, joint pain, choking, blindness, seizures, unconsciousness, and even permanent neurological defects such as paraplegia.

French physiologist Paul Bert (1833-1886) first identified the bends in 1878, and in 1907, John Scott Haldane (1860-1936) developed a method for counteracting decompression sickness. He calculated a set of decompression tables that advised limits for the amount of time at given depths. He recommended what he called stage decompression, which means that the ascending diver stops every few feet during ascension and waits for a few minutes at each level, allowing the body tissues time to adjust to the new pressure. Modern divers use a decompression chamber, a sealed container that simulates the stages of decompression.

BATHYSPHERE, SCUBA, AND BATHYSCAPHE. In 1930, the American naturalist William Beebe (1877-1962) and American engineer Otis Barton created the bathysphere. This was the first submersible that provided the divers inside with adequate protection from external pressure. Made of steel and spherical in shape, the bathysphere had thick quartz windows and was capable of maintaining ordinary atmosphere pressure even when lowered by a cable to relatively great depths. In 1934, a bathysphere descended to what was then an extremely impressive depth: 3,028 ft (923 m). However, the bathysphere was difficult to operate and maneuver, and in time it was be replaced by a more workable vessel, the bathyscaphe.

Before the bathyscaphe appeared, however, in 1943, two Frenchmen created a means for divers to descend without the need for any sort of external chamber. Certainly a diver with this new apparatus could not go to anywhere near the same depths as those approached by the bathysphere; nonetheless, the new aqualung made it possible to spend an extended time under the surface without need for air. It was now theoretically feasible for a diver to go below without any need for help or supplies from above, because he carried his entire oxygen supply on his back. The name of one of inventors, Emile Gagnan, is hardly a household word; but that of the other—Jacques Cousteau (1910-1997)—certainly is. So, too, is the name of their invention: the *s*elf-*c*ontained *u*nderwater *b*reathing *a*pparatus, better known as scuba.

The most important feature of the scuba gear was the demand regulator, which made it possible for the divers to breathe air at the same pressure as their underwater surroundings. This in turn facilitated breathing in a more normal, comfortable manner. Another important feature of a modern diver's equipment is a buoyancy compensation device. Like a ship atop the water, a diver wants to have only so much buoyancy—not so much that it causes him to surface.

As for the bathyscaphe—a term whose two Greek roots mean "deep" and "boat"—it made its debut five years after scuba gear. Built by the Swiss physicist and adventurer Auguste Piccard (1884-1962), the bathyscaphe consisted of two compartments: a heavy steel crew cabin that was resistant to sea pressure, and above it, a larger, light container called a float. The float was filled with gasoline, which in this case was not used as fuel, but to provide extra buoyancy, because of the gasoline's low specific gravity.

When descending, the occupants of the bathyscaphe—there could only be two, since the pressurized chamber was just 79 in (2.01 m) in diameter—released part of the gasoline to decrease buoyancy. They also carried iron ballast pellets on board, and these they released when preparing to ascend. Thanks to battery-driven screw propellers, the bathyscaphe was much more maneuverable than the bathysphere had ever been; furthermore, it was designed to reach depths that Beebe and Barton could hardly have conceived.

REACHING NEW DEPTHS. It took several years of unsuccessful dives, but in 1953 a bathyscaphe set the first of many depth records. This first craft was the *Trieste,* manned by Piccard and his son Jacques, which descended 10,335 ft (3,150 m) below the Mediterranean, off Capri, Italy. A year later, in the Atlantic Ocean off Dakar, French West Africa (now Senegal), French divers Georges Houot and Pierre-Henri Willm reached 13,287 ft (4,063 m) in the *FNRS 3.*

Then in 1960, Jacques Piccard and United States Navy Lieutenant Don Walsh set a record that still stands: 35,797 ft (10,911 m)—23% greater than the height of Mt. Everest, the world's tallest peak. This they did in the *Trieste* some 250 mi (402 km) southeast of Guam at the Mariana Trench, the deepest spot in the Pacific Ocean and indeed the deepest spot on Earth. Piccard and Walsh went all the way to the bottom, a descent that took them 4 hours, 48 minutes. Coming up took 3 hours, 17 minutes.

Thirty-five years later, in 1995, the Japanese craft *Kaiko* also made the Mariana descent and confirmed the measurements of Piccard and Walsh. But the achievement of the *Kaiko* was not nearly as impressive of that of the *Trieste's* two-man crew: the *Kaiko,* in fact, had no crew. By the 1990s, sophisticated remote-sensing technology had made it possible to send down unmanned ocean expeditions, and it became less necessary to expose human beings to the incredible risks encountered by the Piccards, Walsh, and others.

FILMING TITANIC. An example of such an unmanned vessel is the one featured in the opening minutes of the Academy Award-winning motion picture *Titanic* (1997). The vessel itself, whose sinking in 1912 claimed more than 1,000 lives, rests at such a great depth in the North Atlantic that it is impractical either to raise it, or to send manned expeditions to explore the interior of the wreck. The best solution, then, is a remotely operated vessel of the kind also used for purposes such as mapping the ocean floor, exploring for petroleum and other deposits, and gathering underwater plate technology data.

The craft used in the film, which has "arms" for grasping objects, is of a variety specially designed for recovering items from shipwrecks. For the scenes that showed what was supposed to be the *Titanic* as an active vessel, director James Cameron used a 90% scale model that depicted the ship's starboard side—the side hit by the iceberg. Therefore, when showing its port side, as when it was leaving the Southampton, England, dock on April 15, 1912, all shots had to be reversed: the actual signs on the dock were in reverse lettering in order to appear correct when seen in the final version. But for scenes of the wrecked vessel lying at the bottom of the ocean, Cameron used the real *Titanic.*

To do this, he had to use a submersible; but he did not want to shoot only from inside the submersible, as had been done in the 1992 IMAX film *Titanica.* Therefore, his brother Mike Cameron, in cooperation with Panavision, built a special camera that could withstand 400 atm ($3.923 \cdot 10^7$ Pa)—that is, 400 times the air pressure at sea level. The camera was attached to the outside of the submersible, which for these external shots was manned by Russian submarine operators.

Because the special camera only held twelve minutes' worth of film, it was necessary to make a total of twelve dives. On the last two, a remotely operated submersible entered the wreck, which would have been too dangerous for the humans in the manned craft. Cameron had intended the remotely operated submersible as a mere prop, but in the end its view inside the ruined *Titanic* added one of the most poignant touches in the entire film. To these he later added scenes involving objects specific to the film's plot, such as the safe. These he shot in a controlled underwater environment designed to look like the interior of the *Titanic.*

THE DIVERS PICTURED HERE HAVE ASCENDED FROM A SUNKEN SHIP AND HAVE STOPPED AT THE 10-FT (3-M) DECOMPRESSION LEVEL TO AVOID GETTING DECOMPRESSION SICKNESS, BETTER KNOWN AS THE "BENDS." *(Photograph, copyright Jonathan Blair/Corbis. Reproduced by permission.)*

Into the Skies

In the earlier description of Piccard's bathyscaphe design, it was noted that the craft consisted of two compartments: a heavy steel crew cabin resistant to sea pressure, and above it a larger, light container called a float. If this sounds rather like the structure of a hot-air balloon, there is no accident in that.

In 1931, nearly two decades before the bathyscaphe made its debut, Piccard and another Swiss scientist, Paul Kipfer, set a record of a different kind with a balloon. Instead of going lower than anyone ever had, as Piccard and his son Jacques did in 1953—and as Jacques and Walsh did in an even greater way in 1960—Piccard and Kipfer went higher than ever, ascending to 55,563 ft (16,940 m). This made them the first two men to penetrate the stratosphere, which is the next atmospheric layer above the troposphere, a layer approximately 10 mi (16.1 km) high that covers the surface of Earth.

Piccard, without a doubt, experienced the greatest terrestrial altitude range of any human being over a lifetime: almost 12.5 mi (20.1 km) from his highest high to his lowest low, 84% of it above sea level and the rest below. His career, then, was a tribute to the power of buoyant force—and to the power of overcoming buoyant force for the purpose of descending to the ocean depths. Indeed, the same can be said of the Piccard family as a whole: not only did Jacques set the world's depth record, but years later, Jacques's son Bertrand took to the skies for another record-setting balloon flight.

In 1999, Bertrand Piccard and British balloon instructor Brian Wilson became the first men to circumnavigate the globe in a balloon, the *Breitling Orbiter 3.* The craft extended 180 ft (54.86) from the top of the envelope—the part of the balloon holding buoyant gases—to the bottom of the gondola, the part holding riders. The pressurized cabin had one bunk in which one pilot could sleep while the other flew, and up front was a computerized control panel which allowed the pilot to operate the burners, switch propane tanks, and release empty ones. It took Piccard and Wilson just 20 days to circle the Earth—a far cry from the first days of ballooning two centuries earlier.

THE FIRST BALLOONS. The Piccard family, though Swiss, are francophone; that is, they come from the French-speaking part of Switzerland. This is interesting, because the history of human encounters with buoyancy—below the ocean and even more so in the air—has been heavily dominated by French names. In fact, it was the French brothers, Joseph-Michel (1740-1810) and Jacques-Etienne (1745-1799) Montgolfier, who launched the first balloon in 1783. These two became to balloon flight what two other brothers, the Americans Orville and Wilbur Wright, became with regard to the invention that superseded the balloon twelve decades later: the airplane.

On that first flight, the Montgolfiers sent up a model 30 ft (9.15 m) in diameter, made of linen-lined paper. It reached a height of 6,000 ft (1,828 m), and stayed in the air for 10 minutes before coming back down. Later that year, the Montgolfiers sent up the first balloon flight with living creatures—a sheep, a rooster, and a duck—and still later in 1783, Jean-François Pilatre de Rozier (1756-1785) became the first human being to ascend in a balloon.

Rozier only went up 84 ft (26 m), which was the length of the rope that tethered him to the ground. As the makers and users of balloons learned how to use ballast properly, however, flight times were extended, and balloon flight became ever more practical. In fact, the world's first military use of flight dates not to the twentieth century but to the eighteenth—1794, specifically, when France created a balloon corps.

HOW A BALLOON FLOATS. There are only three gases practical for lifting a balloon: hydrogen, helium, and hot air. Each is much less than dense than ordinary air, and this gives them their buoyancy. In fact, hydrogen is the lightest gas known, and because it is cheap to produce, it would be ideal—except for the fact that it is extremely flammable. After the 1937 crash of the airship *Hindenburg,* the era of hydrogen use for lighter-than-air transport effectively ended.

Helium, on the other hand, is perfectly safe and only slightly less buoyant than hydrogen. This makes it ideal for balloons of the sort that children enjoy at parties; but helium is expensive, and therefore impractical for large balloons. Hence, hot air—specifically, air heated to a temperature of about 570°F (299°C), is the only truly viable option.

Charles's law, one of the laws regarding the behavior of gases, states that heating a gas will increase its volume. Gas molecules, unlike their liquid or solid counterparts, are highly non-attractive—that is, they tend to spread toward relatively great distances from one another. There is already a great deal of empty space between gas molecules, and the increase in volume only increases the amount of empty space. Hence, density is lowered, and the balloon floats.

AIRSHIPS. Around the same time the Montgolfier brothers launched their first balloons, another French designer, Jean-Baptiste-Marie Meusnier, began experimenting with a

Once considered obsolete, blimps are enjoying a renaissance among scientists and government agencies. The blimp pictured here, the Aerostat blimp, is equipped with radar for drug enforcement and instruments for weather observation. *(Corbis. Reproduced by permission.)*

more streamlined, maneuverable model. Early balloons, after all, could only be maneuvered along one axis, up and down: when it came to moving sideways or forward and backward, they were largely at the mercy of the elements.

It was more than a century before Meusnier's idea—the prototype for an airship—became a reality. In 1898, Alberto Santos-Dumont of Brazil combined a balloon with a propeller powered by an internal-combustion instrument, creating a machine that improved on the balloon, much as the bathyscaphe later improved on the bathysphere. Santos-Dumont's airship was non-rigid, like a balloon. It also used hydrogen, which is apt to contract during descent and collapse the envelope. To counter this problem, Santos-Dumont created the ballonet, an internal airbag designed to provide buoyancy and stabilize flight.

One of the greatest figures in the history of lighter-than-air flight—a man whose name, along with blimp and dirigible, became a synonym for the airship—was Count Ferdinand von Zeppelin (1838-1917). It was he who created a lightweight structure of aluminum girders and rings that made it possible for an airship to remain rigid under varying atmospheric conditions. Yet Zeppelin's earliest launches, in the decade that followed 1898, were fraught with a number of problems—not least of which were disasters caused by the flammability of hydrogen.

Zeppelin was finally successful in launching airships for public transport in 1911, and the quarter-century that followed marked the golden age of airship travel. Not that all was "golden" about this age: in World War I, Germany used airships as bombers, launching the first London blitz in May 1915. By the time Nazi Germany initiated the more famous World War II London blitz 25 years later, ground-based anti-aircraft technology would have made quick work of any zeppelin; but by then, airplanes had long since replaced airships.

During the 1920s, though, airships such as the *Graf Zeppelin* competed with airplanes as a mode of civilian transport. It is a hallmark of the perceived safety of airships over airplanes at the time that in 1928, the *Graf Zeppelin* made its first transatlantic flight carrying a load of passengers. Just a year earlier, Charles Lindbergh had made the first-ever solo, nonstop transatlantic flight in an airplane. Today this would be the equivalent of someone flying to the Moon, or perhaps even Mars, and there was no question of carrying pas-

KEY TERMS

ARCHIMEDES'S PRINCIPLE: A rule of physics which holds that the buoyant force of an object immersed in fluid is equal to the weight of the fluid displaced by the object. It is named after the Greek mathematician, physicist, and inventor Archimedes (c. 287-212 B.C.), who first identified it.

BALLAST: A heavy substance that, by increasing the weight of an object experiencing buoyancy, improves its stability.

BUOYANCY: The tendency of an object immersed in a fluid to float. This can be explained by Archimedes's principle.

DENSITY: Mass divided by volume.

DISPLACEMENT: A measure of the weight of the fluid that has had to be moved out of position so that an object can be immersed. If a ship set down in the ocean causes 1,000 tons of water to be displaced, it is said to possess a displacement of 1,000 tons.

FLUID: Any substance, whether gas or liquid, that conforms to the shape of its container.

FORCE: The product of mass multiplied by acceleration.

MASS: A measure of inertia, indicating the resistance of an object to a change in its motion. For an object immerse in fluid, mass is equal to volume multiplied by density.

PRESSURE: The exertion of force over a two-dimensional area; hence the formula for pressure is force divided by area. The British system of measures typically reckons pressure in pounds per square inch. In metric terms, this is measured in terms of newtons (N) per square meter, a figure known as a pascal (Pa.)

SPECIFIC GRAVITY: The density of an object or substance relative to the density of water; or more generally, the ratio between the densities of two objects or substances.

VOLUME: The amount of three-dimensional space occupied by an object. Volume is usually measured in cubic units.

WEIGHT: A force equal to mass multiplied by the acceleration due to gravity (32 ft/9.8 m/sec^2). For an object immersed in fluid, weight is the same as volume multiplied by density multiplied by gravitational acceleration.

sengers. Furthermore, Lindbergh was celebrated as a hero for the rest of his life, whereas the passengers aboard the *Graf Zeppelin* earned no more distinction for bravery than would pleasure-seekers aboard a cruise.

THE LIMITATIONS OF LIGHTER-THAN-AIR TRANSPORT. For a few years, airships constituted the luxury liners of the skies; but the *Hindenburg* crash signaled the end of relatively widespread airship transport. In any case, by the time of the 1937 *Hindenburg* crash, lighter-than-air transport was no longer the leading contender in the realm of flight technology.

Ironically enough, by 1937 the airplane had long since proved itself more viable—even though it was actually heavier than air. The principles that make an airplane fly have little to do with buoyancy as such, and involve differences in pressure rather than differences in density. Yet the replacement of lighter-than-air craft on the cutting edge of flight did not mean that balloons and airships were relegated to the museum; instead, their purposes changed.

The airship enjoyed a brief resurgence of interest during World War II, though purely as a surveillance craft for the United States military. In the period after the war, the U.S. Navy hired the Goodyear Tire and Rubber Company to produce airships, and as a result of this relationship Goodyear created the most visible airship since the *Graf Zeppelin* and the *Hindenburg:* the Goodyear Blimp.

BLIMPS AND BALLOONS: ON THE CUTTING EDGE?. The blimp, known to viewers of countless sporting events, is much better-suited than a plane or helicopter to providing TV cameras with an aerial view of a stadium—and advertisers with a prominent billboard. Military forces and science communities have also found airships useful for unexpected purposes. Their virtual invisibility with regard to radar has reinvigorated interest in blimps on the part of the U.S. Department of Defense, which has discussed plans to use airships as radar platforms in a larger Strategic Air Initiative. In addition, French scientists have used airships for studying rain forest treetops or canopies.

Balloons have played a role in aiding space exploration, which is emblematic of the relationship between lighter-than-air transport and more advanced means of flight. In 1961, Malcolm D. Ross and Victor A. Prother of the U.S. Navy set the balloon altitude record with a height of 113,740 ft (34,668 m.) The technology that enabled their survival at more than 21 mi (33.8 km) in the air was later used in creating life-support systems for astronauts.

Balloon astronomy provides some of the clearest images of the cosmos: telescopes mounted on huge, unmanned balloons at elevations as high as 120,000 ft (35,000 m)—far above the dust and smoke of Earth—offer high-resolution images. Balloons have even been used on other planets: for 46 hours in 1985, two balloons launched by the unmanned Soviet expedition to Venus collected data from the atmosphere of that planet.

American scientists have also considered a combination of a large hot-air balloon and a smaller helium-filled balloon for gathering data on the surface and atmosphere of Mars during expeditions to that planet. As the air balloon is heated by the Sun's warmth during the day, it would ascend to collect information on the atmosphere. (In fact the "air" heated would be from the atmosphere of Mars, which is composed primarily of carbon dioxide.) Then at night when Mars cools, the air balloon would lose its buoyancy and descend, but the helium balloon would keep it upright while it collected data from the ground.

WHERE TO LEARN MORE

"*Buoyancy*" (Web site). <http://www.aquaholic.com/gasses/laws.htm> (March 12, 2001).

"*Buoyancy*" (Web site). <http://www.uncwil.edu/nurc/aquarius/lessons/buoyancy.htm> (March 12, 2001).

"*Buoyancy Basics*" *Nova/PBS* (Web site). <http://www.pbs.org/wgbh/nova/lasalle/buoybasics.html> (March 12, 2001).

Challoner, Jack. *Floating and Sinking.* Austin, TX: Raintree Steck-Vaughn, 1997.

Cobb, Allan B. *Super Science Projects About Oceans.* New York: Rosen, 2000.

Gibson, Gary. *Making Things Float and Sink.* Illustrated by Tony Kenyon. Brookfield, CT: Copper Beeck Brooks, 1995.

Taylor, Barbara. *Liquid and Buoyancy.* New York: Warwick Press, 1990.

STATICS

STATICS AND EQUILIBRIUM

PRESSURE

ELASTICITY

STATICS AND EQUILIBRIUM

CONCEPT

Statics, as its name suggests, is the study of bodies at rest. Those bodies may be acted upon by a variety of forces, but as long as the lines of force meet at a common point and their vector sum is equal to zero, the body itself is said to be in a state of equilibrium. Among the topics of significance in the realm of statics is center of gravity, which is relatively easy to calculate for simple bodies, but much more of a challenge where aircraft or ships are concerned. Statics is also applied in analysis of stress on materials—from a picture frame to a skyscraper.

HOW IT WORKS

Equilibrium and Vectors

Essential to calculations in statics is the use of vectors, or quantities that have both magnitude and direction. By contrast, a scalar has only magnitude. If one says that a certain piece of property has an area of one acre, there is no directional component. Nor is there a directional component involved in the act of moving the distance of 1 mi (1.6 km), since no statement has been made as to the direction of that mile. On the other hand, if someone or something experiences a displacement, or change in position, of 1 mi to the northeast, then what was a scalar description has been placed in the language of vectors.

Not only are mass and speed (as opposed to velocity) considered scalars; so too is time. This might seem odd at first glance, but—on Earth at least, and outside any special circumstances posed by quantum mechanics—time can only move forward. Hence, direction is not a factor. By contrast, force, equal to mass multiplied by acceleration, is a vector. So too is weight, a specific type of force equal to mass multiplied by the acceleration due to gravity (32 ft or [9.8 m] / sec^2). Force may be in any direction, but the direction of weight is always downward along a vertical plane.

VECTOR SUMS. Adding scalars is simple, since it involves mere arithmetic. The addition of vectors is more challenging, and usually requires drawing a diagram, for instance, if trying to obtain a vector sum for the velocity of a car that has maintained a uniform speed, but has changed direction several times.

One would begin by representing each vector as an arrow on a graph, with the tail of each vector at the head of the previous one. It would then be possible to draw a vector from the tail of the first to the head of the last. This is the sum of the vectors, known as a resultant, which measures the net change.

Suppose, for instance, that a car travels north 5 mi (8 km), east 2 mi (3.2 km), north 3 mi (4.8 km), east 3 mi, and finally south 3 mi. One must calculate its net displacement—in other words, not the sum of all the miles it has traveled, but the distance and direction between its starting point and its end point. First, one draws the vectors on a piece of graph paper, using a logical system that treats the y axis as the north-south plane, and the x axis as the east-west plane. Each vector should be in the form of an arrow pointing in the appropriate direction.

Having drawn all the vectors, the only remaining one is between the point where the car's journey ends and the starting point—that is, the resultant. The number of sides to the resulting shape is always one more than the number of vectors being added; the final side is the resultant.

In this particular case, the answer is fairly easy. Because the car traveled north 5 mi and ultimately moved east by 5 mi, returning to a position of 5 mi north, the segment from the resultant forms the hypotenuse of an equilateral (that is, all sides equal) right triangle. By applying the Pythagorean theorem, which states that the square of the length of the hypotenuse is equal to the sum of the squares of the other two sides, one quickly arrives at a figure of 7.07 m (11.4 km) in a northeasterly direction. This is the car's net displacement.

Calculating Force and Tension in Equilibrium

Using vector sums, it is possible to make a number of calculations for objects in equilibrium, but these calculations are somewhat more challenging than those in the car illustration. One form of equilibrium calculation involves finding tension, or the force exerted by a supporting object on an object in equilibrium—a force that is always equal to the amount of weight supported. (Another way of saying this is that if the tension on the supporting object is equal to the weight it supports, then the supported object is in equilibrium.)

In calculations for tension, it is best to treat the supporting object—whether it be a rope, picture hook, horizontal strut or some other item—as though it were weightless. One should begin by drawing a free-body diagram, a sketch showing all the forces acting on the supported object. It is not necessary to show any forces (other than weight) that the object itself exerts, since those do not contribute to its equilibrium.

RESOLVING X AND Y COMPONENTS. As with the distance vector graph discussed above, next one must equate these forces to the x and y axes. The distance graph example involved only segments already parallel to x and y, but suppose—using the numbers already discussed—the graph had called for the car to move in a perfect 45°-angle to the northeast along a distance of 7.07 mi. It would then have been easy to resolve this distance into an x component (5 mi east) and a y component (5 mi north)—which are equal to the other two sides of the equilateral triangle.

This resolution of x and y components is more challenging for calculations involving equilibrium, but once one understands the principle involved, it is easy to apply. For example, imagine a box suspended by two ropes, neither of which is at a 90°-angle to the box. Instead, each rope is at an acute angle, rather like two segments of a chain holding up a sign.

The x component will always be the product of tension (that is, weight) multiplied by the cosine of the angle. In a right triangle, one angle is always equal to 90°, and thus by definition, the other two angles are acute, or less than 90°. The angle of either rope is acute, and in fact, the rope itself may be considered the hypotenuse of an imaginary triangle. The base of the triangle is the x axis, and the angle between the base and the hypotenuse is the one under consideration.

Hence, we have the use of the cosine, which is the ratio between the adjacent leg (the base) of the triangle and the hypotenuse. Regardless of the size of the triangle, this figure is a constant for any particular angle. Likewise, to calculate the y component of the angle, one uses the sine, or the ratio between the opposite side and the hypotenuse. Keep in mind, once again, that the adjacent leg for the angle is by definition the same as the x axis, just as the opposite leg is the same as the y axis. The cosine (abbreviated cos), then, gives the x component of the angle, as the sine (abbreviated sin) does the y component.

REAL-LIFE APPLICATIONS

Equilibrium and Center of Gravity in Real Objects

Before applying the concept of vector sums to matters involving equilibrium, it is first necessary to clarify the nature of equilibrium itself—what it is and what it is not. Earlier it was stated that an object is in equilibrium if the vector sum of the forces acting on it are equal to zero—as long as those forces meet at a common point.

This is an important stipulation, because it is possible to have lines of force that cancel one another out, but nonetheless cause an object to move. If a force of a certain magnitude is applied to the right side of an object, and a line of force of equal magnitude meets it exactly from the left, then the object is in equilibrium. But if the line of force from the right is applied to the top of the object, and the line of force from the left to the bottom, then they do not meet at a common point, and the object is not in equilibrium. Instead, it is experiencing torque, which will cause it to rotate.

VARIETIES OF EQUILIBRIUM. There are two basic conditions of equilibrium. The term "translational equilibrium" describes an object that experiences no linear (straight-line) acceleration; on the other hand, an object experiencing no rotational acceleration (a component of torque) is said to be in rotational equilibrium.

Typically, an object at rest in a stable situation experiences both linear and rotational equilibrium. But equilibrium itself is not necessarily stable. An empty glass sitting on a table is in stable equilibrium: if it were tipped over slightly—that is, with a force below a certain threshold—then it would return to its original position. This is true of a glass sitting either upright or upside-down.

Now imagine if the glass were somehow propped along the edge of a book sitting on the table, so that the bottom of the glass formed the hypotenuse of a triangle with the table as its base and the edge of the book as its other side. The glass is in equilibrium now, but unstable equilibrium, meaning that a slight disturbance—a force from which it could recover in a stable situation—would cause it to tip over.

If, on the other hand, the glass were lying on its side, then it would be in a state of neutral equilibrium. In this situation, the application of force alongside the glass will not disturb its equilibrium. The glass will not attempt to seek stable equilibrium, nor will it become more unstable; rather, all other things being equal, it will remain neutral.

CENTER OF GRAVITY. Center of gravity is the point in an object at which the weight below is equal to the weight above, the weight in front equal to the weight behind, and the weight to the left equal to the weight on the right. Every object has just one center of gravity, and if the object is suspended from that point, it will not rotate.

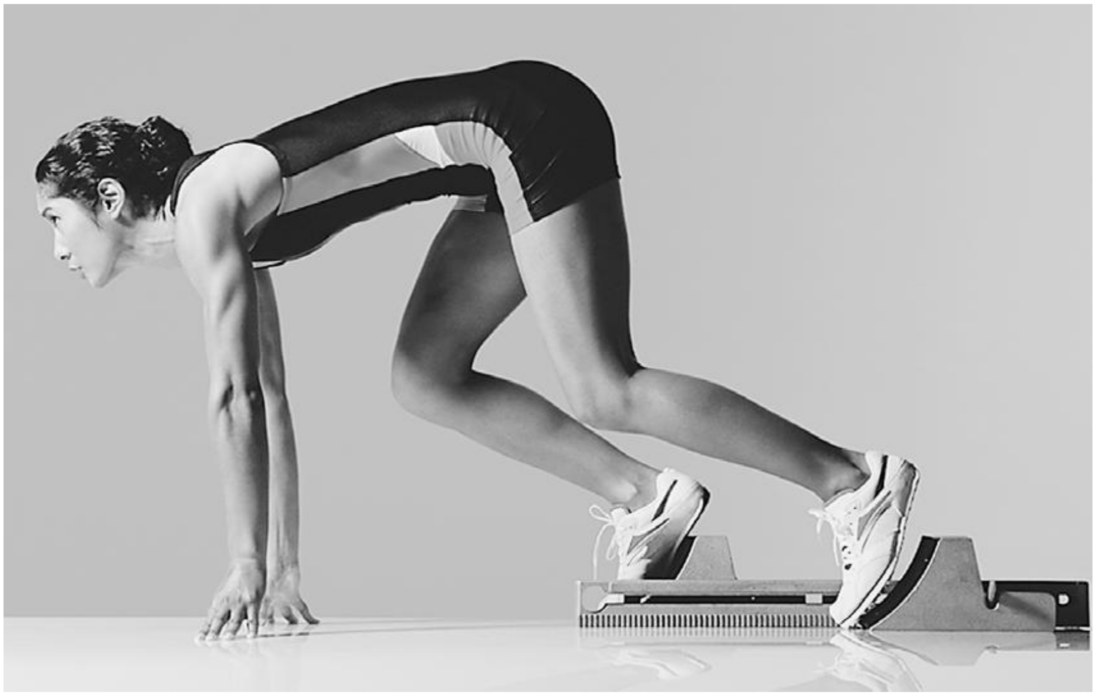
One interesting aspect of an object's center of gravity is that it does not necessarily have to be within the object itself. When a swimmer is poised in a diving stance, as just before the starting bell in an Olympic competition, the swimmer's center of gravity is to the front—some distance from his or her chest. This is appropriate, since the objective is to get into the water as quickly as possible once the race starts.

A glass sitting on a table is in a state of stable equilibrium. *(Photograph by John Wilkes Studio/Corbis. Reproduced by permission.)*

By contrast, a sprinter's stance places the center of gravity well within the body, or at least firmly surrounded by the body—specifically, at the place where the sprinter's rib cage touches the forward knee. This, too, fits with the needs of the athlete in the split-second following the starting gun. The sprinter needs to have as much traction as possible to shoot forward, rather than forward and downward, as the swimmer does.

Tension Calculations

In the earlier discussion regarding the method of calculating tension in equilibrium, two of the three steps of this process were given: first, draw a free-body diagram, and second, resolve the forces into x and y components. The third step is to set the force components along each axis equal to zero—since, if the object is truly in equilibrium, the sum of forces will indeed equal zero. This makes it possible, finally, to solve the equations for the net tension.



IN THE STARTING BLOCKS, A SPRINTER'S CENTER OF GRAVITY IS ALIGNED ALONG THE RIB CAGE AND FORWARD KNEE, THUS MAXIMIZING THE RUNNER'S ABILITY TO SHOOT FORWARD OUT OF THE BLOCKS. *(Photograph by Ronnen Eshel/Corbis. Reproduced by permission.)*

Imagine a picture that weighs 100 lb (445 N) suspended by a wire, the left side of which may be called segment *A*, and the right side segment *B*. The wire itself is not perfectly centered on the picture-hook: *A* is at a 30° angle, and *B* on a 45° angle. It is now possible to find the tension on both.

First, one can resolve the horizontal components by the formula $F_x = T_{Bx} + T_{Ax} = 0$, meaning that the *x* component of force is equal to the product of tension for the *x* component of *B*, added to the product of tension for the *x* component of *A*, which in turn is equal to zero. Given the 30°-angle of *A*, $_{Ax} = 0.866$, which is the cosine of 30°. $_{Bx}$ is equal to cos 45°, which equals 0.707. (Recall the earlier discussion of distance, in which a square with sides 5 mi long was described: its hypotenuse was 7.07 mi, and 5/7.07 = 0.707.)

Because *A* goes off to the left from the point at which the picture is attached to the wire, this places it on the negative portion of the x axis. Therefore, the formula can now be restated as $T_B(0.707)–T_A(0.866) = 0$. Solving for T_B reveals that it is equal to $T_A(0.866/0.707) = (1.22)T_A$. This will be substituted for T_B in the formula for the total force along the y component.

However, the y-force formula is somewhat different than for x: since weight is exerted along the y axis, it must be subtracted. Thus, the formula for the y component of force is $F_y = T_{Ay} + T_{By}–w = 0$. (Note that the y components of both *A* and *B* are positive: by definition, this must be so for an object suspended from some height.)

Substituting the value for T_B obtained above, $(1.22)T_A$, makes it possible to complete the equation. Since the sine of 30° is 0.5, and the sine of 45° is 0.707—the same value as its cosine—one can state the equation thus: $T_A(0.5) + (1.22)T_A(0.707)–100$ lb = 0. This can be restated as $T_A(0.5 + (1.22 \cdot 0.707)) = T_A(1.36) = 100$ lb. Hence, $T_A = (100 \text{ lb}/1.36) = 73.53$ lb. Since $T_B = (1.22)T_A$, this yields a value of 89.71 lb for T_B.

Note that T_A and T_B actually add up to considerably more than 100 lb. This, however, is known as an algebraic sum—which is very similar to an arithmetic sum, inasmuch as algebra is simply a generalization of arithmetic. What is important here, however, is the vector sum, and the vector sum of T_A and T_B equals 100 lb.

CALCULATING CENTER OF GRAVITY. Rather than go through another lengthy calculation for center of gravity, we will explain the principles behind such calculations.

It is easy to calculate the center of gravity for a regular shape, such as a cube or sphere—assuming, of course, that the mass and therefore the weight is evenly distributed throughout the object. In such a case, the center of gravity is the geometric center. For an irregular object, however, center of gravity must be calculated.

An analogy regarding United States demographics may help to highlight the difference between geometric center and center of gravity. The geographic center of the U.S., which is analogous to geometric center, is located near the town of Castle Rock in Butte County, South Dakota. (Because Alaska and Hawaii are so far west of the other 48 states—and Alaska, with its great geographic area, is far to the north—the data is skewed in a northwestward direction. The geographic center of the 48 contiguous states is near Lebanon, in Smith County, Kansas.)

The geographic center, like the geometric center of an object, constitutes a sort of balance point in terms of physical area: there is as much U.S. land to the north of Castle Rock, South Dakota, as to the south, and as much to the east as to the west. The population center is more like the center of gravity, because it is a measure, in some sense, of "weight" rather than of volume—though in this case concentration of people is substituted for concentration of weight. Put another way, the population center is the balance point of the population, if it were assumed that every person weighed the same amount.

Naturally, the population center has been shifting westward ever since the first U.S. census in 1790, but it is still skewed toward the east: though there is far more U.S. land west of the Mississippi River, there are still more people east of it. Hence, according to the 1990 U.S. census, the geographic center is some 1,040 mi (1,664 km) in a southeastward direction from the population center: just northwest of Steelville, Missouri, and a few miles west of the Mississippi.

The United States, obviously, is an "irregular object," and calculations for either its geographic or its population center represent the mastery of numerous mathematical principles. How, then, does one find the center of gravity for a much smaller irregular object? There are a number of methods, all rather complex.

To measure center of gravity in purely physical terms, there are a variety of techniques relating to the shape of the object to be measured. There is also a mathematical formula, which involves first treating the object as a conglomeration of several more easily measured objects. Then the x components for the mass of each "sub-object" can be added, and divided by the combined mass of the object as a whole. The same can be done for the y components.

USING EQUILIBRIUM CALCULATIONS

One reason for making center of gravity calculations is to ensure that the net force on an object passes through that center. If it does not, the object will start to rotate—and for an airplane, for instance, this could be disastrous. Hence, the builders and operators of aircraft make exceedingly detailed, complicated calculations regarding center of gravity. The same is true for shipbuilders and shipping lines: if a ship's center of gravity is not vertically aligned with the focal point of the buoyant force exerted on it by the water, it may well sink.

In the case of ships and airplanes, the shapes are so irregular that center of gravity calculations require intensive analyses of the many components. Hence, a number of companies that supply measurement equipment to the aerospace and maritime industries offer center of gravity measurement instruments that enable engineers to make the necessary calculations.

On dry ground, calculations regarding equilibrium are likewise quite literally a life and death matter. In the earlier illustration, the object in equilibrium was merely a picture hanging from a wire—but what if it were a bridge or a building? The results of inaccurate estimates of net force could affect the lives of many people. Hence, structural engineers make detailed analyses of stress, once again using series of calculations that make the picture-frame illustration above look like the simplest of all arithmetic problems.

WHERE TO LEARN MORE

Beiser, Arthur. *Physics,* 5th ed. Reading, MA: Addison-Wesley, 1991.

"Determining Center of Gravity" National Aeronautics and Space Administration (Web site). <http://www.grc.nasa.gov/WWW/K-12/airplane/cg.html> (March 19, 2001).

KEY TERMS

ACCELERATION: A change in velocity.

CENTER OF GRAVITY: The point on an object at which the total weights on either side of all axes (x, y, and z) are identical. Each object has just one center of gravity, and if it is suspended from that point, it will be in a state of perfect rotational equilibrium.

COSINE: For an acute (less than 90°) angle in a right triangle, the cosine (abbreviated cos) is the ratio between the adjacent leg and the hypotenuse. Regardless of the size of the triangle, this figure is a constant for any particular angle.

DISPLACEMENT: Change in position.

EQUILIBRIUM: A state in which vector sum for all lines of force on an object is equal to zero. An object that experiences no linear acceleration is said to be in translational equilibrium, and one that experiences no rotational acceleration is referred to as being in rotational equilibrium. An object may also be in stable, unstable, or neutral equilibrium.

FORCE: The product of mass multiplied by acceleration.

FREE-BODY DIAGRAM: A sketch showing all the outside forces acting on an object in equilibrium.

HYPOTENUSE: In a right triangle, the side opposite the right angle.

RESULTANT: The sum of two or more vectors, which measures the net change in distance and direction.

RIGHT TRIANGLE: A triangle that includes a right (90°) angle. The other two angles are, by definition, acute, or less than 90°.

SCALAR: A quantity that possesses only magnitude, with no specific direction. Mass, time, and speed are all scalars. The opposite of a scalar is a vector.

SINE: For an acute (less than 90°) angle in a right triangle, the sine (abbreviated sin) is the ratio between the opposite leg and the hypotenuse. Regardless of the size of the triangle, this figure is a constant for any particular angle.

STATICS: The study of bodies at rest. Those bodies may be acted upon by a variety of forces, but as long as the vector sum for all those lines of force is equal to zero, the body itself is said to be in a state of equilibrium.

TENSION: The force exerted by a supporting object on an object in equilibrium—a force that is always equal to the amount of weight supported.

VECTOR: A quantity that possesses both magnitude and direction. Force is a vector; so too is acceleration, a component of force; and likewise weight, a variety of force. The opposite of a vector is a scalar.

VECTOR SUM: A calculation, made by different methods according to the factor being analyzed—for instance, velocity or force—that yields the net result of all the vectors applied in a particular situation.

VELOCITY: The speed of an object in a particular direction. Velocity is thus a vector quantity.

WEIGHT: A measure of the gravitational force on an object; the product of mass multiplied by the acceleration due to gravity.

"*Equilibrium and Statics*" (Web site). <http://www.glenbrook.k12.il.us/gbssci/phys/Class/vectors/u313c.html> (March 19, 2001).

"*Exploratorium Snack: Center of Gravity.*" *The Exploratorium* (Web site). <http://www.exploratorium.edu/snacks/center_of_gravity.html> (March 19, 2001).

Faivre d'Arcier, Marima. *What Is Balance?* Illustrated by Volker Theinhardt. New York: Viking Kestrel, 1986.

Taylor, Barbara. *Weight and Balance.* Photographs by Peter Millard. New York: F. Watts, 1990.

"*Where Is Your Center of Gravity?*" *The K-8 Aeronautics Internet Textbook* (Web site). <http://wing.ucdavis.edu/Curriculums/Forces_Motion/center_howto.html> (March 19, 2001).

Wood, Robert W. *Mechanics Fundamentals.* Illustrated by Bill Wright. Philadelphia, PA: Chelsea House, 1997.

Zubrowski, Bernie. *Mobiles: Building and Experimenting with Balancing Toys.* Illustrated by Roy Doty. New York: Morrow Junior Books, 1993.

PRESSURE

CONCEPT

Pressure is the ratio of force to the surface area over which it is exerted. Though solids exert pressure, the most interesting examples of pressure involve fluids—that is, gases and liquids—and in particular water and air. Pressure plays a number of important roles in daily life, among them its function in the operation of pumps and hydraulic presses. The maintenance of ordinary air pressure is essential to human health and well-being: the body is perfectly suited to the ordinary pressure of the atmosphere, and if that pressure is altered significantly, a person may experience harmful or even fatal side-effects.

HOW IT WORKS

FORCE AND SURFACE AREA

When a force is applied perpendicular to a surface area, it exerts pressure on that surface equal to the ratio of F to A, where F is the force and A the surface area. Hence, the formula for pressure (p) is $p = F/A$. One interesting consequence of this ratio is the fact that pressure can increase or decrease without any change in force—in other words, if the surface becomes smaller, the pressure becomes larger, and vice versa.

If one cheerleader were holding another cheerleader on her shoulders, with the girl above standing on the shoulder blades of the girl below, the upper girl's feet would exert a certain pressure on the shoulders of the lower girl. This pressure would be equal to the upper girl's weight (F, which in this case is her mass multiplied by the downward acceleration due to gravity) divided by the surface area of her feet. Suppose, then, that the upper girl executes a challenging acrobatic move, bringing her left foot up to rest against her right knee, so that her right foot alone exerts the full force of her weight. Now the surface area on which the force is exerted has been reduced to half its magnitude, and thus the pressure on the lower girl's shoulder is twice as great.

For the same reason—that is, that reduction of surface area increases net pressure—a well-delivered karate chop is much more effective than an open-handed slap. If one were to slap a board squarely with one's palm, the only likely result would be a severe stinging pain on the hand. But if instead one delivered a blow to the board, with the hand held perpendicular—provided, of course, one were an expert in karate—the board could be split in two. In the first instance, the area of force exertion is large and the net pressure to the board relatively small, whereas in the case of the karate chop, the surface area is much smaller—and hence, the pressure is much larger.

Sometimes, a greater surface area is preferable. Thus, snowshoes are much more effective for walking in snow than ordinary shoes or boots. Ordinary footwear is not much larger than the surface of one's foot, perfectly appropriate for walking on pavement or grass. But with deep snow, this relatively small surface area increases the pressure on the snow, and causes one's feet to sink. The snowshoe, because it has a surface area significantly larger than that of a regular shoe, reduces the ratio of force to surface area and therefore, lowers the net pressure.

The same principle applies with snow skis and water skis. Like a snowshoe, a ski makes it possible for the skier to stay on the surface of the snow, but unlike a snowshoe, a ski is long and thin, thus enabling the skier to glide more effectively down a snow-covered hill. As for skiing on water, people who are experienced at this sport can ski barefoot, but it is tricky. Most beginners require water skis, which once again reduce the net pressure exerted by the skier's weight on the surface of the water.

MEASURING PRESSURE

Pressure is measured by a number of units in the English and metric—or, as it is called in the scientific community, SI—systems. Because $p = F/A$, all units of pressure represent some ratio of force to surface area. The principle SI unit is called a pascal (Pa), or 1 N/m^2. A newton (N), the SI unit of force, is equal to the force required to accelerate 1 kilogram of mass at a rate of 1 meter per second squared. Thus, a Pascal is equal to the pressure of 1 newton over a surface area of 1 square meter.

In the English or British system, pressure is measured in terms of pounds per square inch, abbreviated as $lbs./in^2$. This is equal to $6.89 \cdot 10^3$ Pa, or 6,890 Pa. Scientists—even those in the United States, where the British system of units prevails—prefer to use SI units. However, the British unit of pressure is a familiar part of an American driver's daily life, because tire pressure in the United States is usually reckoned in terms of pounds per square inch. (The recommended tire pressure for a mid-sized car is typically 30-35 lb/in^2.)

Another important measure of pressure is the atmosphere (atm), which the average pressure exerted by air at sea level. In English units, this is equal to 14.7 $lbs./in^2$, and in SI units to $1.013 \cdot 10^5$ Pa—that is, 101,300 Pa. There are also two other specialized units of pressure measurement in the SI system: the bar, equal to 10^5 Pa, and the torr, equal to 133 Pa. Meteorologists, scientists who study weather patterns, use the millibar (mb), which, as its name implies, is equal to 0.001 bars. At sea level, atmospheric pressure is approximately 1,013 mb.

THE BAROMETER. The torr, once known as the "millimeter of mercury," is equal to the pressure required to raise a column of mercury (chemical symbol Hg) 1 mm. It is named for the Italian physicist Evangelista Torricelli (1608-1647), who invented the barometer, an instrument for measuring atmospheric pressure.

IN THE INSTANCE OF ONE CHEERLEADER STANDING ON ANOTHER'S SHOULDERS, THE CHEERLEADER'S FEET EXERT DOWNWARD PRESSURE ON HER PARTNER'S SHOULDERS. THE PRESSURE IS EQUAL TO THE GIRL'S WEIGHT DIVIDED BY THE SURFACE AREA OF HER FEET. *(Photograph by James L. Amos/Corbis. Reproduced by permission.)*

The barometer, constructed by Torricelli in 1643, consisted of a long glass tube filled with mercury. The tube was open at one end, and turned upside down into a dish containing more mercury: hence, the open end was submerged in mercury while the closed end at the top constituted a vacuum—that is, an area in which the pressure is much lower than 1 atm.

The pressure of the surrounding air pushed down on the surface of the mercury in the bowl, while the vacuum at the top of the tube provided an area of virtually no pressure, into which the mercury could rise. Thus, the height to which the mercury rose in the glass tube represented normal air pressure (that is, 1 atm.) Torricelli discovered that at standard atmospheric pressure, the column of mercury rose to 760 millimeters.

The value of 1 atm was thus established as equal to the pressure exerted on a column of mercury 760 mm high at a temperature of 0°C (32°F). Furthermore, Torricelli's invention eventually became a fixture both of scientific labora-

The air pressure on top of Mount Everest, the world's tallest peak, is very low, making breathing difficult. Most climbers who attempt to scale Everest thus carry oxygen tanks with them. Shown here is Jim Whittaker, the first American to climb Everest. *(Photograph by Galen Rowell/Corbis. Reproduced by permission.)*

tories and of households. Since changes in atmospheric pressure have an effect on weather patterns, many home indoor-outdoor thermometers today also include a barometer.

Pressure and Fluids

In terms of physics, both gases and liquids are referred to as fluids—that is, substances that conform to the shape of their container. Air pressure and water pressure are thus specific subjects under the larger heading of "fluid pressure." A fluid responds to pressure quite differently than a solid does. The density of a solid makes it resistant to small applications of pressure, but if the pressure increases, it experiences tension and, ultimately, deformation. In the case of a fluid, however, stress causes it to flow rather than to deform.

There are three significant characteristics of the pressure exerted on fluids by a container. First of all, a fluid in a container experiencing no external motion exerts a force perpendicular to the walls of the container. Likewise, the container walls exert a force on the fluid, and in both cases, the force is always perpendicular to the walls.

In each of these three characteristics, it is assumed that the container is finite: in other words, the fluid has nowhere else to go. Hence, the second statement: the external pressure exerted on the fluid is transmitted uniformly. Note that the preceding statement was qualified by the term "external": the fluid itself exerts pressure whose force component is equal to its weight. Therefore, the fluid on the bottom has much greater pressure than the fluid on the top, due to the weight of the fluid above it.

Third, the pressure on any small surface of the fluid is the same, regardless of that surface's orientation. In other words, an area of fluid perpendicular to the container walls experiences the same pressure as one parallel or at an angle to the walls. This may seem to contradict the first principle, that the force is perpendicular to the walls of the container. In fact, force is a vector quantity, meaning that it has both magnitude and direction, whereas pressure is a scalar, meaning that it has magnitude but no specific direction.

REAL-LIFE APPLICATIONS

Pascal's Principle and the Hydraulic Press

The three characteristics of fluid pressure described above have a number of implications and applications, among them, what is known as Pascal's principle. Like the SI unit of pressure, Pascal's principle is named after Blaise Pascal (1623-1662), a French mathematician and physicist who formulated the second of the three statements: that the external pressure applied on a fluid is transmitted uniformly throughout the entire body of that fluid. Pascal's principle became the basis for one of the important machines ever developed, the hydraulic press.

A simple hydraulic press of the variety used to raise a car in an auto shop typically consists of two large cylinders side by side. Each cylinder contains a piston, and the cylinders are connected at the bottom by a channel containing fluid. Valves control flow between the two cylinders. When one applies force by pressing down the piston in one cylinder (the input cylinder), this yields a uniform pressure that causes output in the second cylinder, pushing up a piston that raises the car.

In accordance with Pascal's principle, the pressure throughout the hydraulic press is the same, and will always be equal to the ratio between force and pressure. As long as that ratio is the same, the values of *F* and *A* may vary. In the case of an auto-shop car jack, the input cylinder has a relatively small surface area, and thus, the amount of force that must be applied is relatively small as well. The output cylinder has a relatively large surface area, and therefore, exerts a relatively large force to lift the car. This, combined with the height differential between the two cylinders (discussed in the context of mechanical advantage elsewhere in this book), makes it possible to lift a heavy automobile with a relatively small amount of effort.

THE HYDRAULIC RAM. The car jack is a simple model of the hydraulic press in operation, but in fact, Pascal's principle has many more applications. Among these is the hydraulic ram, used in machines ranging from bulldozers to the hydraulic lifts used by firefighters and utility workers to reach heights. In a hydraulic ram, however, the characteristics of the input and output cylinders are reversed from those of a car jack.

The input cylinder, called the master cylinder, has a large surface area, whereas the output cylinder (called the slave cylinder) has a small surface area. In addition—though again, this is a factor related to mechanical advantage rather than pressure, per se—the master cylinder is short, whereas the slave cylinder is tall. Owing to the larger surface area of the master cylinder compared to that of the slave cylinder, the hydraulic ram is not considered efficient in terms of mechanical advantage: in other words, the force input is much greater than the force output.

Nonetheless, the hydraulic ram is as well-suited to its purpose as a car jack. Whereas the jack is made for lifting a heavy automobile through a short vertical distance, the hydraulic ram carries a much lighter cargo (usually just one person) through a much greater vertical range—to the top of a tree or building, for instance.

EXPLOITING PRESSURE DIFFERENCES

PUMPS. A pump utilizes Pascal's principle, but instead of holding fluid in a single container, a pump allows the fluid to escape. Specifically, the pump utilizes a pressure difference, causing the fluid to move from an area of higher pressure to one of lower pressure. A very simple example of this is a siphon hose, used to draw petroleum from a car's gas tank. Sucking on one end of the hose creates an area of low pressure compared to the relatively high-pressure area of the gas tank. Eventually, the gasoline will come out of the low-pressure end of the hose. (And with luck, the person siphoning will be able to anticipate this, so that he does not get a mouthful of gasoline!)

The piston pump, more complex, but still fairly basic, consists of a vertical cylinder along which a piston rises and falls. Near the bottom of the cylinder are two valves, an inlet valve through which fluid flows into the cylinder, and an outlet valve through which fluid flows out of it. On the suction stroke, as the piston moves upward, the inlet valve opens and allows fluid to enter the cylinder. On the downstroke, the inlet valve closes while the outlet valve opens, and the pressure provided by the piston on the fluid forces it through the outlet valve.

One of the most obvious applications of the piston pump is in the engine of an automobile. In this case, of course, the fluid being pumped is gasoline, which pushes the pistons by providing a series of controlled explosions created by the spark plug's ignition of the gas. In another variety of piston pump—the kind used to inflate a basketball or a bicycle tire—air is the fluid being pumped. Then there is a pump for water, which pumps drinking water from the ground It may also be used to remove desirable water from an area where it is a hindrance, for instance, in the bottom of a boat.

BERNOULLI'S PRINCIPLE. Though Pascal provided valuable understanding with regard to the use of pressure for performing work, the thinker who first formulated general principles regarding the relationship between fluids and pressure was the Swiss mathematician and physicist Daniel Bernoulli (1700-1782). Bernoulli is considered the father of fluid mechanics, the study of the behavior of gases and liquids at rest and in motion.

While conducting experiments with liquids, Bernoulli observed that when the diameter of a pipe is reduced, the water flows faster. This suggested to him that some force must be acting upon the water, a force that he reasoned must arise from differences in pressure. Specifically, the slower-moving fluid in the wider area of pipe had a greater pressure than the portion of the fluid moving through the narrower part of the pipe. As a result, he concluded that pressure and velocity are inversely related—in other words, as one increases, the other decreases.

Hence, he formulated Bernoulli's principle, which states that for all changes in movement, the sum of static and dynamic pressure in a fluid remain the same. A fluid at rest exerts static pressure, which is commonly meant by "pressure," as in "water pressure." As the fluid begins to move, however, a portion of the static pressure—proportional to the speed of the fluid—is converted to what is known as dynamic pressure, or the pressure of movement. In a cylindrical pipe, static pressure is exerted perpendicular to the surface of the container, whereas dynamic pressure is parallel to it.

According to Bernoulli's principle, the greater the velocity of flow in a fluid, the greater the dynamic pressure and the less the static pressure: in other words, slower-moving fluid exerts greater pressure than faster-moving fluid. The discovery of this principle ultimately made possible the development of the airplane.

As fluid moves from a wider pipe to a narrower one, the volume of that fluid that moves a given distance in a given time period does not change. But since the width of the narrower pipe is smaller, the fluid must move faster (that is, with greater dynamic pressure) in order to move the same amount of fluid the same distance in the same amount of time. One way to illustrate this is to observe the behavior of a river: in a wide, unconstricted region, it flows slowly, but if its flow is narrowed by canyon walls, then it speeds up dramatically.

Bernoulli's principle ultimately became the basis for the airfoil, the design of an airplane's wing when seen from the end. An airfoil is shaped like an asymmetrical teardrop laid on its side, with the "fat" end toward the airflow. As air hits the front of the airfoil, the airstream divides, part of it passing over the wing and part passing under. The upper surface of the airfoil is curved, however, whereas the lower surface is much straighter.

As a result, the air flowing over the top has a greater distance to cover than the air flowing under the wing. Since fluids have a tendency to compensate for all objects with which they come into contact, the air at the top will flow faster to meet with air at the bottom at the rear end of the wing. Faster airflow, as demonstrated by Bernoulli, indicates lower pressure, meaning that the pressure on the bottom of the wing keeps the airplane aloft.

Buoyancy and Pressure

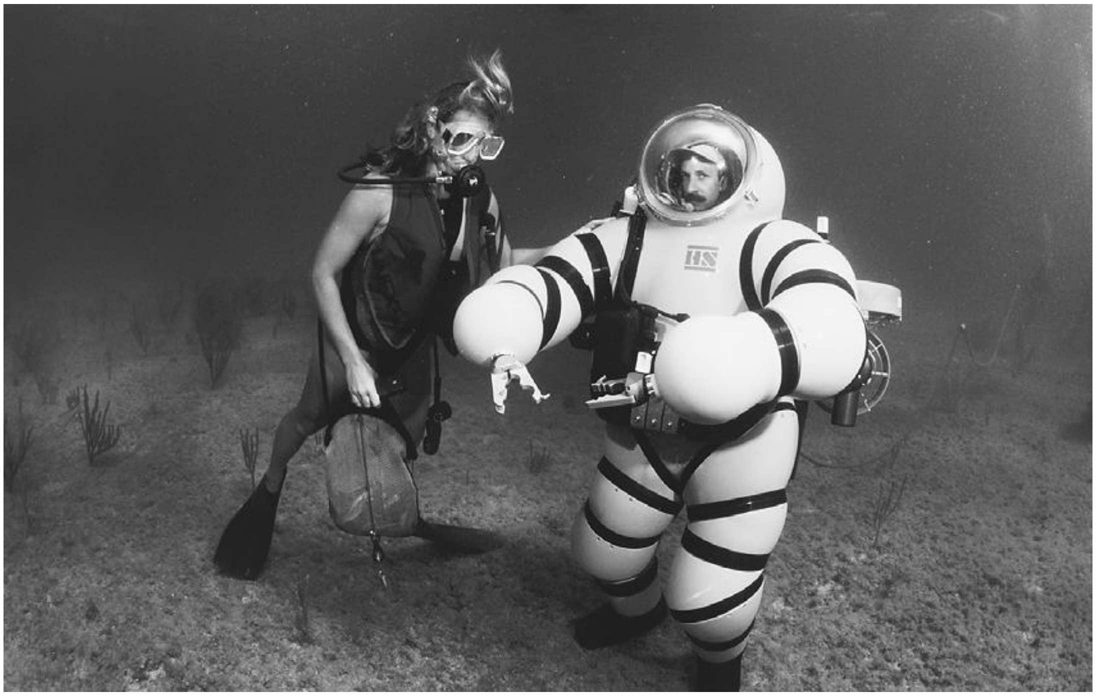
One hundred and twenty years before the first successful airplane flight by the Wright brothers in 1903, another pair of brothers—the Montgolfiers of France—developed another means of flight. This was the balloon, which relied on an entirely different principle to get off the ground: buoyancy, or the tendency of an object immersed in a fluid to float. As with Bernoulli's principle, however, the concept of buoyancy is related to pressure.

In the third century B.C., the Greek mathematician, physicist, and inventor Archimedes (c. 287-212 B.C.) discovered what came to be known as Archimedes's principle, which holds that the buoyant force of an object immersed in fluid is equal to the weight of the fluid displaced by the object. This is the reason why ships float: because the buoyant, or lifting, force of them is less than equal to the weight of the water they displace.

The hull of a ship is designed to displace or move a quantity of water whose weight is greater than that of the vessel itself. The weight of the displaced water—that is, its mass multiplied by the downward acceleration caused by gravity—is equal to the buoyant force that the ocean exerts on the ship. If the ship weighs less than the water it displaces, it will float; but if it weighs more, it will sink.

The factors involved in Archimedes's principle depend on density, gravity, and depth rather than pressure. However, the greater the depth within a fluid, the greater the pressure that pushes against an object immersed in the fluid. Moreover, the overall pressure at a given depth in a fluid is related in part to both density and gravity, components of buoyant force.

PRESSURE AND DEPTH. The pressure that a fluid exerts on the bottom of its container is equal to *dgh,* where *d* is density, *g* the acceleration due to gravity, and *h* the depth of the container. For any portion of the fluid, *h* is equal to its depth within the container, meaning that



This yellow diving suit, called a "newt suit," is specially designed to withstand the enormous water pressure that exists at lower depths of the ocean. *(Photograph by Amos Nachoum/Corbis. Reproduced by permission.)*

the deeper one goes, the greater the pressure. Furthermore, the total pressure within the fluid is equal to $dgh + p_{external}$, where $p_{external}$ is the pressure exerted on the surface of the fluid. In a piston-and-cylinder assembly, this pressure comes from the piston, but in water, the pressure comes from the atmosphere.

In this context, the ocean may be viewed as a type of "container." At its surface, the air exerts downward pressure equal to 1 atm. The density of the water itself is uniform, as is the downward acceleration due to gravity; the only variable, then, is *h,* or the distance below the surface. At the deepest reaches of the ocean, the pressure is incredibly great—far more than any human being could endure. This vast amount of pressure pushes upward, resisting the downward pressure of objects on its surface. At the same time, if a boat's weight is dispersed properly along its hull, the ship maximizes area and minimizes force, thus exerting a downward pressure on the surface of the water that is less than the upward pressure of the water itself. Hence, it floats.

Pressure and the Human Body

AIR PRESSURE. The Montgolfiers used the principle of buoyancy not to float on the water, but to float in the sky with a craft lighter than air. The particulars of this achievement are discussed elsewhere, in the context of buoyancy; but the topic of lighter-than-air flight suggests another concept that has been alluded to several times throughout this essay: air pressure.

Just as water pressure is greatest at the bottom of the ocean, air pressure is greatest at the surface of the Earth—which, in fact, is at the bottom of an "ocean" of air. Both air and water pressure are examples of hydrostatic pressure—the pressure that exists at any place in a body of fluid due to the weight of the fluid above. In the case of air pressure, air is pulled downward by the force of Earth's gravitation, and air along the surface has greater pressure due to the weight (a function of gravity) of the air above it. At great heights above Earth's surface, however, the gravitational force is diminished, and, thus, the air pressure is much smaller.

In ordinary experience, a person's body is subjected to an impressive amount of pressure. Given the value of atmospheric pressure discussed earlier, if one holds out one's hand—assuming that the surface is about 20 in^2 (0.129 m^2)—the force of the air resting on it is nearly 300 lb (136 kg)! How is it, then, that one's

KEY TERMS

ATMOSPHERE: A measure of pressure, abbreviated "atm" and equal to the average pressure exerted by air at sea level. In English units, this is equal to 14.7 pounds per square inch, and in SI units to 101,300 pascals.

BAROMETER: An instrument for measuring atmospheric pressure.

BUOYANCY: The tendency of an object immersed in a fluid to float.

FLUID: Any substance, whether gas or liquid, that conforms to the shape of its container.

FLUID MECHANICS: The study of the behavior of gases and liquids at rest and in motion.

HYDROSTATIC PRESSURE: the pressure that exists at any place in a body of fluid due to the weight of the fluid above.

PASCAL: The principle SI or metric unit of pressure, abbreviated "Pa" and equal to 1 N/m^2.

PASCAL'S PRINCIPLE: A statement, formulated by French mathematician and physicist Blaise Pascal (1623-1662), which holds that the external pressure applied on a fluid is transmitted uniformly throughout the entire body of that fluid.

PRESSURE: The ratio of force to surface area, when force is applied in a direction perpendicular to that surface. The formula for pressure (p) is $p = F/A$, where F is force and A the surface area.

hand is not crushed by all this weight? The reason is that the human body itself is under pressure, and that the interior of the body exerts a pressure equal to that of the air.

THE RESPONSE TO CHANGES IN AIR PRESSURE. The human body is, in fact, suited to the normal air pressure of 1 atm, and if that external pressure is altered, the body undergoes changes that may be harmful or even fatal. A minor example of this is the "popping" in the ears that occurs when one drives through the mountains or rides in an airplane. With changes in altitude come changes in pressure, and thus, the pressure in the ears changes as well.

As noted earlier, at higher altitudes, the air pressure is diminished, which makes it harder to breathe. Because air is a gas, its molecules have a tendency to be non-attractive: in other words, when the pressure is low, they tend to move away from one another, and the result is that a person at a high altitude has difficulty getting enough air into his or her lungs. Runners competing in the 1968 Olympics at Mexico City, a town in the mountains, had to train in high-altitude environments so that they would be able to breathe during competition. For baseball teams competing in Denver, Colorado (known as "the Mile-High City"), this disadvantage in breathing is compensated by the fact that lowered pressure and resistance allows a baseball to move more easily through the air.

If a person is raised in such a high-altitude environment, of course, he or she becomes used to breathing under low air pressure conditions. In the Peruvian Andes, for instance, people spend their whole lives at a height more than twice as great as that of Denver, but a person from a low-altitude area should visit such a locale only after taking precautions. At extremely great heights, of course, no human can breathe: hence airplane cabins are pressurized. Most planes are equipped with oxygen masks, which fall from the ceiling if the interior of the cabin experiences a pressure drop. Without these masks, everyone in the cabin would die.

BLOOD PRESSURE. Another aspect of pressure and the human body is blood pressure. Just as 20/20 vision is ideal, doctors recommend a target blood pressure of "120 over 80"—but what does that mean? When a person's blood pressure is measured, an inflatable cuff is wrapped around the upper arm at the same level as the heart. At the same time, a stethoscope is placed along an artery in the lower arm to monitor the sound of the blood flow. The cuff is inflated to stop the blood flow, then the pressure is released until the blood just begins flowing again, producing a gurgling sound in the stethoscope.

The pressure required to stop the blood flow is known as the systolic pressure, which is equal to the maximum pressure produced by the heart. After the pressure on the cuff is reduced until the blood begins flowing normally—which is reflected by the cessation of the gurgling sound in the stethoscope—the pressure of the artery is measured again. This is the diastolic pressure, or the pressure that exists within the artery between strokes of the heart. For a healthy person, systolic pressure should be 120 torr, and diastolic pressure 80 torr.

WHERE TO LEARN MORE

"Atmospheric Pressure: The Force Exerted by the Weight of Air" (Web site). <http://kids.earth.nasa.gov/archive/air_pressure/> (April 7, 2001).

Beiser, Arthur. *Physics,* 5th ed. Reading, MA: Addison-Wesley, 1991.

"Blood Pressure" (Web site). <http://www.mckinley.uiuc.edu/health-info/dis-cond/bloodpr/bloodpr.html> (April 7, 2001).

Clark, John Owen Edward. *The Atmosphere.* New York: Gloucester Press, 1992.

Cobb, Allan B. *Super Science Projects About Oceans.* New York: Rosen, 2000.

"The Physics of Underwater Diving: Pressure Lesson" (Web site). <http://www.uncwil.edu/nurc/aquarius/lessons/pressure.html> (April 7, 2001).

Provenzo, Eugene F. and Asterie Baker Provenzo. *47 Easy-to-Do Classic Experiments.* Illustrations by Peter A. Zorn, Jr. New York: Dover Publications, 1989.

"Understanding Air Pressure" USA Today (Web site). <http://www.usatoday.com/weather/wbarocx.html> (April 7, 2001).

Zubrowski, Bernie. *Balloons: Building and Experimenting with Inflatable Toys.* Illustrated by Roy Doty. New York: Morrow Junior Books, 1990.

ELASTICITY

CONCEPT

Unlike fluids, solids do not respond to outside force by flowing or easily compressing. The term elasticity refers to the manner in which solids respond to stress, or the application of force over a given unit area. An understanding of elasticity—a concept that carries with it a rather extensive vocabulary of key terms—helps to illuminate the properties of objects from steel bars to rubber bands to human bones.

HOW IT WORKS

Characteristics of a Solid

A number of parameters distinguish solids from fluids, a term that in physics includes both gases and liquids. Solids possess a definite volume and a definite shape, whereas gases have neither; liquids have no definite shape.

At the molecular level, particles of solids tend to be precise in their arrangement and close to one another. Liquid molecules are close in proximity (though not as much so as solid molecules), and their arrangement is random, while gas molecules are both random in arrangement and far removed in proximity. Gas molecules are extremely fast-moving, and exert little or no attraction toward one another. Liquid molecules move at moderate speeds and exert a moderate attraction, but solid particles are slow-moving, and have a strong attraction to one another.

One of several factors that distinguishes solids from fluids is their relative response to pressure. Gases tend to be highly compressible, meaning that they respond well to pressure. Liquids tend to be noncompressible, yet because of their fluid characteristics, they experience external pressure uniformly. If one applies pressure to a quantity of water in a closed container, the pressure is equal everywhere in the water. By contrast, if one places a champagne glass upright in a vise and applies pressure until it breaks, chances are that the stem or the base of the glass will be unaffected, because the pressure is not distributed equally throughout the glass.

If the surface of a solid is disturbed, it will resist, and if the force of the disturbance is sufficiently strong, it will deform—for instance, when a steel plate begins to bend under pressure. This deformation will be permanent if the force is powerful enough, as in the above example of the glass in a vise. By contrast, when the surface of a fluid is disturbed, it tends to flow.

Types of Stress

Deformation occurs as a result of stress, whether that stress be in the form of tension, compression, or shear. Tension occurs when equal and opposite forces are exerted along the ends of an object. These operate on the same line of action, but away from each other, thus stretching the object. A perfect example of an object under tension is a rope in the middle of a tug-of-war competition. The adjectival form of "tension" is "tensile": hence the term "tensile stress," which will be discussed later.

Earlier, stress was defined as the application of force over a given unit area, and in fact, the formula for stress can be written as F/A, where F is force and A area. This is also the formula for pressure, though in order for an object to be under pressure, the force must be applied in a direction perpendicular to—and in the same direction as—its surface. The one form of stress that clearly matches these parameters is compression, produced by the action of equal and opposite forces, whose effect is to reduce the length of a material. Thus compression (for example, crushing an aluminum can in one's hand) is both a form of stress and a form of pressure.

Note that compression was defined as reducing length, yet the example given involved a reduction in what most people would call the "width" or diameter of the aluminum can. In fact, width and height are the same as length, for the purposes of most discussions in physics. Length is, along with time, mass, and electric current, one of the fundamental units of measure used to express virtually all other physical quantities. Width and height are simply length expressed in terms of other planes, and within the subject of elasticity, it is not important to distinguish between these varieties of length. (By contrast, when discussing gravitational attraction—which is always vertical—it is obviously necessary to distinguish between "vertical length," or height, and horizontal length.)

The third variety of stress is shear, which occurs when a solid is subjected to equal and opposite forces that do not act along the same line, and which are parallel to the surface area of the object. If a thick hardbound book is lying flat, and a person places a finger on the spine and pushes the front cover away from the spine so that the covers and pages no longer constitute parallel planes, this is an example of shear. Stress resulting from shear is called shearing stress.

HOOKE'S LAW AND ELASTIC LIMIT

To sum up the three varieties of stress, tension stretches an object, compression shrinks it, and shear twists it. In each case, the object is deformed to some degree. This deformation is expressed in terms of strain, or the ratio between change in dimension and the original dimensions of the object. The formula for strain is $\delta L/L_o$, where δL is the change in length (δ, the Greek letter delta, means "change" in scientific notation) and L_o the original length.

Hooke's law, formulated by English physicist Robert Hooke (1635-1703), relates strain to stress. Hooke's law can be stated in simple terms as "the strain is proportional to the stress," and can also be expressed in a formula, $F = ks$, where F is the applied force, s, the resulting change in dimension, and k, a constant whose value is related to the nature and size of the object under stress. The harder the material, the higher the value of k; furthermore, the value of k is directly proportional to the object's cross-sectional area or thickness.

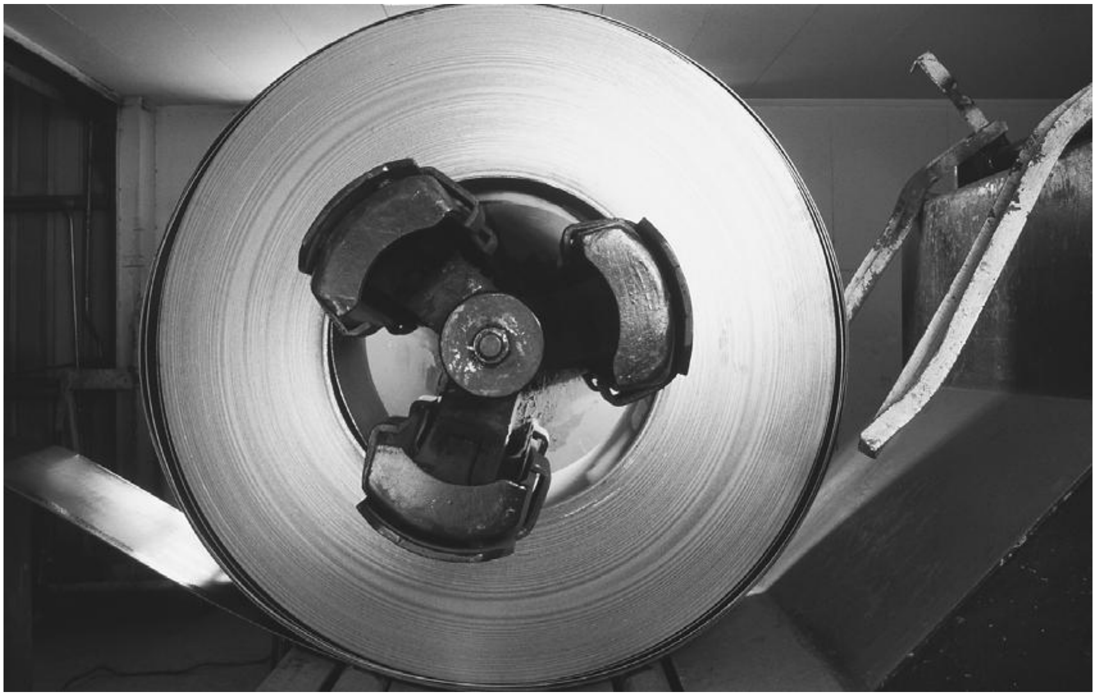
The elastic limit of a given solid is the maximum stress to which it can be subjected without experiencing permanent deformation. Elastic limit will be discussed in the context of several examples below; for now, it is important merely to know that Hooke's law is applicable only as long as the material in question has not reached its elastic limit. The same is true for any modulus of elasticity, or the ratio between a particular type of applied stress and the strain that results. (The term "modulus," whose plural is "moduli," is Latin for "small measure.")

MODULI OF ELASTICITY

In cases of tension or compression, the modulus of elasticity is Young's modulus. Named after English physicist Thomas Young (1773-1829), Young's modulus is simply the ratio between F/A and $\delta L/L_o$—in other words, stress divided by strain. There are also moduli describing the behavior of objects exposed to shearing stress (shear modulus), and of objects exposed to compressive stress from all sides (bulk modulus).

Shear modulus is the relationship of shearing stress to shearing strain. This can be expressed as the ratio between F/A and ϕ. The latter symbol, the Greek letter phi, stands for the angle of shear—that is, the angle of deformation along the sides of an object exposed to shearing stress. The greater the amount of surface area A, the less that surface will be displaced by the force F. On the other hand, the greater the amount of force in proportion to A, the greater the value of ϕ, which measures the strain of an object exposed to shearing stress. (The value of ϕ, however, will usually be well below 90°, and certainly cannot exceed that magnitude.)

With tensile and compressive stress, A is a surface perpendicular to the direction of applied force, but with shearing stress, A is parallel to F. Consider again the illustration used above, of a thick hardbound book lying flat. As noted, when one pushes the front cover from the side so that the covers and pages no longer constitute parallel planes, this is an example of shear. If one pulled the spine and the long end of the pages away



THE MACHINE PICTURED HERE ROLLS OVER STEEL IN ORDER TO BEND IT INTO PIPES. BECAUSE OF ITS ELASTIC NATURE, STEEL CAN BE BENT WITHOUT BREAKING. *(Photograph by Vince Streano/Corbis. Reproduced by permission.)*

from one another, that would be tensile stress, whereas if one pushed in on the sides of the pages and spine, that would be compressive stress. Shearing stress, by contrast, would stress only the front cover, which is analogous to *A* for any object under shearing stress.

The third type of elastic modulus is bulk modulus, which occurs when an object is subjected to compression from all sides—that is, volume stress. Bulk modulus is the relationship of volume stress to volume strain, expressed as the ratio between F/A and $\delta V/V_o$, where δV is the change in volume and V_o is the original volume.

REAL-LIFE APPLICATIONS

ELASTIC AND PLASTIC DEFORMATION

As noted earlier, the elastic limit is the maximum stress to which a given solid can be subjected without experiencing permanent deformation, referred to as plastic deformation. Plastic deformation describes a permanent change in shape or size as a result of stress; by contrast, elastic deformation is only a temporary change in dimension.

A classic example of elastic deformation, and indeed, of highly elastic behavior, is a rubber band: it can be deformed to a length many times its original size, but upon release, it returns to its original shape. Examples of plastic deformation, on the other hand, include the bending of a steel rod under tension or the breaking of a glass under compression. Note that in the case of the steel rod, the object is deformed without rupturing—that is, without breaking or reducing to pieces. The breaking of the glass, however, is obviously an instance of rupturing.

METALS AND ELASTICITY

Metals, in fact, exhibit a number of interesting characteristics with regard to elasticity. With the notable exception of cast iron, metals tend to possess a high degree of ductility, or the ability to be deformed beyond their elastic limits without experiencing rupture. Up to a certain point, the ratio of tension to elongation for metals is high: in other words, a high amount of tension produces only a small amount of elongation. Beyond the elastic limit, however, the ratio is much lower: that is, a relatively small amount of tension produces a high degree of elongation.

Because of their ductility, metals are highly malleable, and, therefore, capable of experienc-

Rubber bands, like the ones shown here formed into a ball, are a classic example of elastic deformation. *(Photograph by Matthew Klein/Corbis. Reproduced by permission.)*

ing mechanical deformation through metallurgical processes, such as forging, rolling, and extrusion. Cold extrusion involves the application of high pressure—that is, a high bulk modulus—to a metal without heating it, and is used on materials such as tin, zinc, and copper to change their shape. Hot extrusion, on the other hand, involves heating a metal to a point of extremely high malleability, and then reshaping it. Metals may also be melted for the purposes of casting, or pouring the molten material into a mold.

ULTIMATE STRENGTH. The tension that a material can withstand is called its ultimate strength, and due to their ductile properties, most metals possess a high value of ultimate strength. It is possible, however, for a metal to break down due to repeated cycles of stress that are well below the level necessary to rupture it. This occurs, for instance, in metal machines such as automobile engines that experience a high frequency of stress cycles during operation.

The high ultimate strength of metals, both in tension and compression, makes them useful in a number of structural capacities. Steel has an ultimate compressive strength 25 times as great as concrete, and an ultimate tensile strength 250 times as great. For this reason, when concrete is poured for building bridges or other large structures, steel rods are inserted in the concrete. Called "rebar" (for "reinforced bars"), the steel rods have ridges along them in order to bond more firmly with the concrete as it dries. As a result, reinforced concrete has a much greater ability than plain concrete to withstand tension and compression.

Steel Bars and Rubber Bands Under Stress

CRYSTALLINE MATERIALS. Metals are crystalline materials, meaning that they are composed of solids called crystals. Particles of crystals are highly ordered, with a definite geometric arrangement repeated in all directions, rather like a honeycomb. (It should be noted, however, that the crystals are not necessarily as uniform in size as the "cells" of the honeycomb.) The atoms of a crystal are arranged in orderly rows, bound to one another by strongly attractive forces that act like microscopic springs.

Just as a spring tends to return to its original length, the highly attractive atoms in a steel bar, when it is stretched, tend to restore it to its original dimensions. Likewise, it takes a great deal of force to pull apart the atoms. When the metal is subjected to plastic deformation, the atoms move

A human bone has a greater "ultimate strength" than that of concrete. *(Ecoscene/Corbis. Reproduced by permission.)*

to new positions and form new bonds. The atoms are incapable of forming bonds; however, when the metal has been subjected to stress exceeding its ultimate strength, at that point, the metal breaks.

The crystalline structure of metal influences its behavior under high temperatures. Heat causes atoms to vibrate, and in the case of metals, this means that the "springs" are stretching and compressing. As temperature increases, so do the vibrations, thus increasing the average distance between atoms. For this reason, under extremely high temperature, the elastic modulus of the metal decreases, and the metal becomes less resistant to stress.

POLYMERS AND ELASTOMERS. Rubber is so elastic in behavior that in everyday life, the term "elastic" is most often used for objects containing rubber: the waistband on a pair of underwear, for instance. The long, thin molecules of rubber, which are arranged side-by-side, are called "polymers," and the super-elastic polymers in rubber are called "elastomers." The chemical bonds between the atoms in a polymer are flexible, and tend to rotate, producing kinks along the length of the molecule.

When a piece of rubber is subjected to tension, as, for instance, if one pulls a rubber band by the ends, the kinks and loops in the elastomers straighten. Once the stress is released, however, the elastomers immediately return to their original shape. The more "kinky" the polymers, the higher the elastic modulus, and hence, the more capable the item is of stretching and rebounding.

It is interesting to note that steel and rubber, materials that are obviously quite different, are both useful in part for the same reason: their high elastic modulus when subjected to tension, and their strength under stress. But a rubber band exhibits behaviors under high temperatures that are quite different from that of a metal: when heated, rubber contracts. It does so quite suddenly, in fact, suggesting that the added energy of the heat allows the bonds in the elastomers to begin rotating again, thus restoring the kinked shape of the molecules.

Bones

The tensile strength in bone fibers comes from the protein collagen, while the compressive strength is largely due to the presence of inorganic (non-living) salt crystals. It may be hard to believe, but bone actually has an ultimate strength—both in tension and compression—greater than that of concrete!

The ultimate strength of most materials is rendered in factors of 10^8 N/m_2—that is, 100,000,000 newtons (the metric unit of force) per square meter. For concrete under tensile stress, the ultimate strength is 0.02, whereas for bone, it is 1.3. Under compressive stress, the values are 0.2 and 1.7, respectively. In fact, the ultimate tensile strength of bone is close to that of cast iron (1.7), though the ultimate compressive strength of cast iron (5.5) is much higher than for bone.

Even with these figures, it may be hard to understand how bone can be stronger than concrete, but that is largely because the volume of concrete used in most situations is much greater than the volume of any bone in the body of a human being. By way of explanation, consider a piece of concrete no bigger than a typical bone: under relatively small amounts of stress, it would crumble.

KEY TERMS

ANGLE OF SHEAR: The angle of deformation on the sides of an object exposed to shearing stress. Its symbol is ϕ (the Greek letter phi), and its value will usually be well below 90°.

BULK MODULUS: The modulus of elasticity for a material subjected to compression on all surfaces—that is, volume stress. Bulk modulus is the relationship of volume stress to volume strain, expressed as the ratio between F/A and dV/Vo, where dV is the change in volume and Vo is the original volume.

COMPRESSION: A form of stress produced by the action of equal and opposite forces, whose effect is to reduce the length of a material. Compression is a form of pressure. When compressive stress is applied to all surfaces of a material, this is known as volume stress.

DUCTILITY: A property whereby a material is capable of being deformed far beyond its elastic limit without experiencing rupture—that is, without breaking. Most metals other than cast iron are highly ductile.

ELASTIC DEFORMATION: A temporary change in shape or size experienced by a solid subjected to stress. Elastic deformation is thus less severe than plastic deformation.

ELASTIC LIMIT: The maximum stress to which a given solid can be subjected without experiencing plastic deformation—that is, without being permanently deformed.

ELASTICITY: The response of solids to stress.

HOOKE'S LAW: A principle of elasticity formulated by English physicist Robert Hooke (1635-1703), who discovered that strain is proportional to stress. Hooke's law can be written as a formula, $F = ks$, where F is the applied force, s the resulting change in dimension, and k a constant whose value is related to the nature and size of the object being subjected to stress. Hooke's law applies only when the elastic limit has not been exceeded.

LENGTH: In discussions of elasticity, "length" refers to an object's dimensions on any given plane, thus, it can be used not only to refer to what is called length in everyday language, but also to width or height.

MODULUS OF ELASTICITY: The ratio between a type of applied stress (that is, tension, compression, and shear) and the strain that results in the object to which stress has been applied. Elastic moduli—including Young's modulus, shearing modulus, and bulk modulus—are applicable only as long as the object's elastic limit has not been reached.

PLASTIC DEFORMATION: A permanent change in shape or size experienced by a solid subjected to stress. Plastic deformation is thus more severe than elastic deformation.

PRESSURE: The ratio of force to surface area, when force is applied in a direction perpendicular to, and in the same direction as, that surface.

KEY TERMS CONTINUED

SHEAR: A form of stress resulting from equal and opposite forces that do not act along the same line. If a thick hardbound book is lying flat, and one pushes the front cover from the side so that the covers and pages no longer constitute parallel planes, this is an example of shear.

SHEAR MODULUS: The modulus of elasticity for an object exposed to shearing stress. It is expressed as the ratio between F/A and ϕ, where ϕ (the Greek letter phi) stands for the angle of shear.

STRAIN: The ratio between the change in dimension experienced by an object that has been subjected to stress, and the original dimensions of the object. The formula for strain is dL/L_o, where dL is the change in length and L_o the original length. Hooke's law, as well as the various moduli of elasticity, relates strain to stress.

STRESS: In general terms, stress is any attempt to deform a solid. Types of stress include tension, compression, and shear. More specifically, stress is the ratio of force to unit area, F/A, where F is force and A area. Thus, it is similar to pressure, and indeed, compression is a form of pressure.

TENSION: A form of stress produced by a force which acts to stretch a material. The adjectival form of "tension" is "tensile": hence the terms "tensile stress" and "tensile strain."

ULTIMATE STRENGTH: The tension that a material can withstand without rupturing. Due to their high levels of ductility, most metals have a high value of ultimate strength.

VOLUME STRESS: The stress that occurs in a material when it is subjected to compression from all sides. The modus of elasticity for volume stress is the bulk modulus.

YOUNG'S MODULUS: A modulus of elasticity describing the relationship between stress to strain for objects under either tension or compression. Named after English physicist Thomas Young (1773-1829), Young's modulus is simply the ratio between F/A and $\delta L/L_o$—in other words, stress divided by strain.

WHERE TO LEARN MORE

Beiser, Arthur. *Physics,* 5th ed. Reading, MA: Addison-Wesley, 1991.

"*Dictionary of Metallurgy*" *Steelmill.com: The Polish Steel Industry Directory* (Web site). <http://www.steelmill.com/DICTIONARY/Diction-ary.htm> (April 9, 2001).

"*Engineering Processes.*" *eFunda.com* (Web site). <http://www.efunda.com/processes/processes_home/process.cfm> (April 9, 2001).

Gibson, Gary. *Making Shapes.* Illustrated by Tony Kenyon. Brookfield, CT: Copper Beech Books, 1996.

"*Glossary of Materials Testing Terms*" (Web site). <http://www.instron.com/apps/glossary> (April 9, 2001).

Goodwin, Peter H. *Engineering Projects for Young Scientists.* New York: Franklin Watts, 1987.

Johnston, Tom. *The Forces with You!* Illustrated by Sarah Pooley. Milwaukee, WI: Gareth Stevens Publishing, 1988.

WORK AND ENERGY

MECHANICAL ADVANTAGE AND SIMPLE MACHINES

CONCEPT

When the term machine is mentioned, most people think of complex items such as an automobile, but, in fact, a machine is any device that transmits or modifies force or torque for a specific purpose. Typically, a machine increases either the force of the person operating it—an aspect quantified in terms of mechanical advantage—or it changes the distance or direction across which that force can be operated. Even a humble screw is a machine; so too is a pulley, and so is one of the greatest machines ever invented: the wheel. Virtually all mechanical devices are variations on three basic machines: the lever, the inclined plane, and the hydraulic press. From these three, especially the first two, arose literally hundreds of machines that helped define history, and which still permeate daily life.

HOW IT WORKS

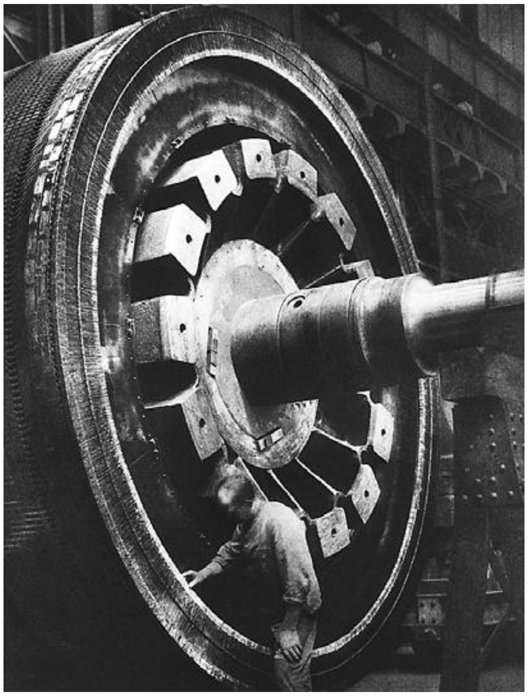
Machines and Classical Mechanics

There are four known types of force in the universe: gravitational, electromagnetic, weak nuclear, and strong nuclear. This was the order in which the forces were identified, and the number of machines that use each force descends in the same order. The essay that follows will make little or no reference to nuclear-powered machines. Somewhat more attention will be paid to electrical machines; however, to trace in detail the development of forces in that context would require a new and somewhat cumbersome vocabulary.

Instead, the machines presented for consideration here depend purely on gravitational force and the types of force explainable purely in a gravitational framework. This is the realm of classical physics, a term used to describe the studies of physicists from the time of Galileo Galilei (1564-1642) to the end of the nineteenth century. During this era, physicists were primarily concerned with large-scale interactions that were easily comprehended by the senses, as opposed to the atomic behaviors that have become the subject of modern physics.

Late in the classical era, the Scottish physicist James Clerk Maxwell (1831-1879)—building on the work of many distinguished predecessors—identified electromagnetic force. For most of the period, however, the focus was on gravitational force and mechanics, or the study of matter, motion, and forces. Likewise, the majority of machines invented and built during most of the classical period worked according to the mechanical principles of plain gravitational force.

This was even true to some extent with the steam engine, first developed late in the seventeenth century and brought to fruition by Scotland's James Watt (1736-1819.) Yet the steam engine, though it involved ordinary mechanical processes in part, represented a new type of machine, which used thermal energy. This is also true of the internal-combustion engine; yet both steam- and gas-powered engines to some extent borrowed the structure of the hydraulic press, one of the three basic types of machine. Then came the development of electronic power, thanks to Thomas Edison (1847-1931) and others, and machines became increasingly divorced from basic mechanical laws.



THE LEVER, LIKE THIS HYDROELECTRIC ENGINE LEVER, IS A SIMPLE MACHINE THAT PERFECTLY ILLUSTRATES THE CONCEPT OF MECHANICAL ADVANTAGE. *(Photograph by E.O. Hoppe/Corbis. Reproduced by permission.)*

The heyday of classical mechanics—when classical studies in mechanics represented the absolute cutting edge of experimentation—was in the period from the beginning of the seventeenth century to the beginning of the nineteenth. One figure held a dominant position in the world of physics during those two centuries, and indeed was the central figure in the history of physics between Galileo and Albert Einstein (1879-1955). This was Sir Isaac Newton (1642-1727), who discerned the most basic laws of physical reality—laws that govern everyday life, including the operation of simple machines.

Newton and his principles are essential to the study that follows, but one other figure deserves "equal billing": the Greek mathematician, physicist, and inventor Archimedes (c. 287-212 B.C.). Nearly 2,000 years before Newton, Archimedes explained and improved a number of basic machines, most notably the lever. Describing the powers of the lever, he is said to have promised, "Give me a lever long enough and a place to stand, and I will move the world." This he demonstrated, according to one story, by moving a fully loaded ship single-handedly with the use of a lever, while remaining seated some distance away.

MECHANICAL ADVANTAGE

A common trait runs through all forms of machinery: mechanical advantage, or the ratio of force output to force input. In the case of the lever, a simple machine that will be discussed in detail below, mechanical advantage is high. In some machines, however, mechanical advantage is actually less than 1, meaning that the resulting force is less than the applied force.

This does not necessarily mean that the machine itself has a flaw; on the contrary, it can mean that the machine has a different purpose than that of a lever. One example of this is the screw: a screw with a high mechanical advantage—that is, one that rewarded the user's input of effort by yielding an equal or greater output—would be useless. In this case, mechanical advantage could only be achieved if the screw backed out from the hole in which it had been placed, and that is clearly not the purpose of a screw.

Here a machine offers an improvement in terms of direction rather than force; likewise with scissors or a fishing rod, both of which will be discussed below, an improvement with regard to distance or range of motion is bought at the expense of force. In these and many more cases, mechanical advantage alone does not measure the benefit. Thus, it is important to keep in mind what was previously stated: a machine either increases force output, or changes the force's distance or direction of operation.

Most machines, however, work best when mechanical advantage is maximized. Yet mechanical advantage—whether in theoretical terms or real-life instances—can only go so high, because there are factors that limit it. For one thing, the operator must give some kind of input to yield an output; furthermore, in most situations friction greatly diminishes output. Hence, in the operation of a car, for instance, one-quarter of the vehicle's energy is expended simply on overcoming the resistance of frictional forces.

For centuries, inventors have dreamed of creating a mechanism with an almost infinite mechanical advantage. This is the much-sought-after "perpetual motion machine," that would only require a certain amount of initial input; after that, the machine would simply run on its own forever. As output compounded over the years, its ratio to input would become so high that the figure for mechanical advantage would approach infinity.

A number of factors, most notably the existence of friction, prevent the perpetual motion machine from becoming anything other than a pipe dream. In outer space, however, the near-absence of friction makes a perpetual motion machine viable: hence, a space probe launched from Earth can travel indefinitely unless or until it enters the gravitational field of some other body in deep space.

The concept of a perpetual motion machine, at least on Earth, is only an idealization; yet idealization does have its place in physics. Physicists discuss most concepts in terms of an idealized state. For instance, when illustrating the acceleration due to gravity experienced by a body in free fall, it is customary to treat such an event as though it were taking place under conditions divorced from reality. To consider the effects of friction, air resistance, and other factors on the body's fall would create an impossibly complicated problem—yet real-world situations are just that complicated.

In light of this tendency to discuss physical processes in idealized terms, it should be noted that there are two types of mechanical advantage: theoretical and actual. Efficiency, as applied to machines in its most specific scientific sense, is the ratio of actual to theoretical mechanical advantage. This in some ways resembles the formula for mechanical advantage itself: once again, what is being measured is the relationship between "output" (the real behavior of the machine) and "input" (the planned behavior of the machine).

As with other mechanical processes, the actual mechanical advantage of a machine is a much more complicated topic than the theoretical mechanical advantage. The gulf between the two, indeed, is enormous. It would be almost impossible to address the actual behavior of machines within an environment framework that includes complexities such as friction.

Each real-world framework—that is, each physical event in the real world—is just a bit different from every other one, due to the many varieties of factors involved. By contrast, the idealized machines of physics problems behave exactly the same way in one imaginary situation after another, assuming outside conditions are the same. Therefore, the only form of mechanical advantage that a physicist can easily discuss is theoretical. For that reason, the term "efficiency" will henceforth be used as a loose synonym for mechanical advantage—even though the technical definition is rather different.

Types of Machines

The term "simple machine" is often used to describe the labor-saving devices known to the ancient world, most of which consisted of only one or two essential parts. Historical sources vary regarding the number of simple machines, but among the items usually listed are levers, pulleys, winches, wheels and axles, inclined planes, wedges, and screws. The list, though long, can actually be reduced to just two items: levers and inclined planes. All the items listed after the lever and before the inclined plane—including the wheel and axle—are merely variations on the lever. The same goes for the wedge and the screw, with regard to the inclined plane.

In fact, all machines are variants on three basic devices: the lever, the inclined plane, and the hydraulic press. Each transmits or modifies force or torque, producing an improvement in force, distance, or direction. The first two, which will receive more attention here, share several aspects not true of the third. First of all, the lever and the inclined plane originated at the beginning of civilization, whereas the hydraulic press is a much more recent invention.

The lever appeared as early as 5000 B.C. in the form of a simple balance scale, and within a few thousand years, workers in the Near East and India were using a crane-like lever called the shaduf to lift containers of water. The shaduf, introduced in Mesopotamia in about 3000 B.C., consisted of a long wooden pole that pivoted on two upright posts. At one end of the lever was a counterweight, and at the other a bucket. The operator pushed down on the pole to fill the bucket with water, and then used the counterweight to assist in lifting the bucket.

The inclined plane made its appearance in the earliest days of civilization, when the Egyptians combined it with rollers in the building of their monumental structures, the pyramids. Modern archaeologists generally believe that Egyptian work gangs raised the huge stone blocks of the pyramids through the use of sloping earthen ramps. These were most probably built up alongside the pyramid itself, and then removed when the structure was completed.

The hydraulic press, like the lift holding up this car in a repair shop, is a relatively recent invention that is used in many modern devices, from car jacks to toilets. *(Photograph by Charles E. Rotkin/Corbis. Reproduced by permission.)*

In contrast to the distant origins of the other two machines, the hydraulic press seems like a mere youngster. Also, the first two clearly arose as practical solutions first: by the time Archimedes achieved a conceptual understanding of these machines, they had been in use a long, long time. The hydraulic press, on the other hand, first emerged from the theoretical studies of the brilliant French mathematician and physicist Blaise Pascal (1623-1662.) Nearly 150 years passed before the English inventor Joseph Bramah (1748-1814) developed a workable hydraulic press in 1796, by which time Watt had already introduced his improved steam engine. The Industrial Revolution was almost underway.

REAL-LIFE APPLICATIONS

The Lever

In its most basic form, the lever consists of a rigid bar supported at one point, known as the fulcrum. One of the simplest examples of a lever is a crowbar, which one might use to move a heavy object, such as a rock. In this instance, the fulcrum could be the ground, though a more rigid "artificial" fulcrum (such as a brick) would probably be more effective.

As the operator of the crowbar pushes down on its long shaft, this constitutes an input of force, variously termed applied force, effort force, or merely effort. Newton's third law of motion shows that there is no such thing as an unpaired force in the universe: every input of force in one area will yield an output somewhere else. In this case, the output is manifested by dislodging the stone—that is, the output force, resistance force, or load.

Use of the lever gives the operator much greater lifting force than that available to a person who tried to lift with only the strength of his or her own body. Like all machines, the lever links input to output, harnessing effort to yield beneficial results—in this case, by translating the input effort into the output effort of a dislodged stone. Note, however, the statement at the beginning of this paragraph: proper use of a lever actually gives a person much greater force than he or she would possess unaided. How can this be?

There is a close relationship between the behavior of the lever and the concept of torque, as, for example, the use of a wrench to remove a lug nut. A wrench, in fact, is a sort of lever (Class

I—a distinction that will be explored below.) In any object experiencing torque, the distance from the pivot point (the lug nut, in this case), to the area where force is being applied is called the moment arm. On the wrench, this is the distance from the lug nut to the place where the operator is pushing on the wrench handle. Torque is the product of force multiplied by moment arm, and the greater the torque, the greater the tendency of the object to be put into rotation. As with machines in general, the greater the input, the greater the output.

The fact that torque is the product of force and moment arm means that if one cannot increase force, it is still possible to gain greater torque by increasing the moment arm. This is the reason why, when one tries and fails to disengage a stubborn lug nut, it is a good idea to get a longer wrench. Likewise with a lever, greater leverage can be gained without applying more force: all one needs is a longer lever arm.

As one might suspect, the lever arm is the distance from the force input to the fulcrum, or from the fulcrum to the force output. If a carpenter is using a nail-puller on the head of a hammer to extract a nail from a board, the lever arm of force input would be from the carpenter's hand gripping the hammer handle to the place where the hammerhead rests against the board. The lever arm of force output would be from the hammerhead to the end of the nail-puller.

With a lever, the input force (that of the carpenter's hand pulling back on the hammer handle) multiplied by the input lever arm is always equal to the output force (that of the nail-puller pulling up the nail) multiplied by the output lever arm. The relationship between input and output force and lever arm then makes it possible to determine a formula for the lever's mechanical advantage.

Since $F_{in}L_{in} = F_{out}L_{out}$ (where F = force and L = lever arm), it is possible to set up an equation for a lever's mechanical advantage. Once again, mechanical advantage is always F_{out}/F_{in}, but with a lever, it is also L_{in}/L_{out}. Hence, the mechanical advantage of a lever is always the same as the inverse ratio of the lever arm. If the input arm is 5 units long and the output arm is 1 unit long, the mechanical advantage will be 5, but if the positions are reversed, it will be 0.2.

CLASSES OF LEVERS. Levers are divided into three classes, depending on the relative positions of the input lever arm, the fulcrum, and the output arm or load. In a Class I lever, such as the crowbar and the wrench, the fulcrum is between the input arm and the output arm. By contrast, a Class II lever, for example, a wheelbarrow, places the output force (the load carried in the barrow itself) between the input force (the action of the operator lifting the handles) and the fulcrum, which in this case is the wheel.

Finally, there is the Class III lever, which is the reverse of a Class II. Here, the input force is between the output force and the fulcrum. The human arm itself is an example of a Class III lever: if one grasps a weight in one's hand, one's bent elbow is the fulcrum, the arm raising the weight is the input force, and the weight held in the hand—now rising—is the output force. The Class III lever has a mechanical advantage of less than 1, but what it loses in force output in gains in range of motion.

The world abounds with levers. Among the Class I varieties in common use are a nail puller on a hammerhead, described earlier, as well as postal scales and pliers. A handcart, though it might seem at first like a wheelbarrow, is actually a Class I lever, because the wheel or fulcrum is between the input effort—the force of a person's hands gripping the handles—and the output, which is the lifting of the load in the handcart itself. Scissors constitute an interesting type of Class I lever, because the force of the output (the cutting blades) is reduced in order to create a greater lever arm for the input, in this case the handles gripped when cutting.

A handheld bottle opener provides an excellent illustration of a Class II lever. Here the fulcrum is on the far end of the opener, away from the operator: the top end of the opener ring, which rests atop the bottle cap. The cap itself is the load, and one provides input force by pulling up on the opener handle, thus prying the cap from the bottle with the lower end of the opener ring. Nail clippers represent a type of combination lever: the handle that one operates is a Class II, while the cutting blades are a Class III.

Whereas Class II levers maximize force at the expense of range of motion, Class III levers operate in exactly the opposite fashion. When using a fishing rod to catch a fish, the fisherman's left hand (assuming he is right-handed) constitutes the fulcrum as it holds the rod just below the reel assembly. The right hand supplies the effort, jerking upward, while the fish is the load. The purpose here is not to raise a heavy object (one reason why a fishing rod may break if one catches too large a fish) but rather to use the increased lever arm for one's advantage in catching an object at some distance. Similarly, a hammer, which constitutes a Class III lever, with the operator's wrist as fulcrum, magnifies the motion of the operator's hand with a hammerhead that cuts a much wider arc.

Many machines that arose in the Industrial Age are a combination of many levers—that is, a compound lever. In a manual typewriter or piano, for instance, each key is a complex assembly of levers designed for a given task. An automobile, too, uses multiple levers—most notably, a special variety known as the wheel and axle.

THE WHEEL AND AXLE. A wheel is a variation on a Class I lever, but it represents such a stunning technological advancement that it deserves to be considered on its own. When driving a car, the driver places input force on the rim of the steering wheel, whose fulcrum is the center of the wheel. The output force is translated along the steering column to the driveshaft.

The combination of wheel and axle overcomes one factor that tends to limit the effectiveness of most levers, regardless of class: limited range of motion. An axle is really a type of wheel, though it has a smaller radius, which means that the output lever arm is correspondingly smaller. Given what was already said about the equation for mechanical advantage in a lever, this presents a very fortunate circumstance.

When a wheel turns, it has a relatively large lever arm (the rim), that turns a relatively small lever arm, the axle. Because the product of input force and input lever arm must equal output force multiplied by output lever arm, this means that the output force will be higher than the input force. Therefore, the larger the wheel in proportion to the axle, the greater the mechanical advantage.

It is for this reason that large vehicles without power steering often have very large steering wheels, which have a larger range of motion and, thus, a greater torque on the axle—that is, the steering column. Some common examples of the wheel-and-axle principle in operation today include a doorknob and a screwdriver; however, long before the development of the wheel and axle, there were wheels alone.

There is nothing obvious about the wheel, and in fact, it is not nearly as old an invention as most people think. Until the last few centuries, most peoples in sub-Saharan Africa, remote parts of central and northern Asia, the Americas, and the Pacific Islands remained unaware of it. This did not necessarily make them "primitive": even the Egyptians who built the Great Pyramid of Cheops in about 2550 B.C. had no concept of the wheel.

What the Egyptians did have, however, were rollers—-most often logs, onto which a heavy object was hoisted using a lever. From rollers developed the idea of a sledge, a sled-like device for sliding large loads atop a set of rollers. A sledge appears in a Sumerian illustration from about 3500 B.C., the oldest known representation of a wheel-like object.

The transformation from the roller-and-sledge assembly to wheeled vehicles is not as easy as it might seem, and historians still disagree as to the connection. Whatever the case, it appears that the first true wheels originated in Sumer (now part of Iraq) in about 3500 B.C. These were tripartite wheels, made by attaching three pieces of wood and then cutting out a circle. This made a much more durable wheel than a sawed-off log, and also overcame the fact that few trees are perfectly round.

During the early period of wheeled transportation, from 3500 to 2000 B.C., donkeys and oxen rather than horses provided the power, in part because wheeled vehicles were not yet made for the speeds that horses could achieve. Hence, it was a watershed event when wheelmakers began fashioning axles as machines separate from the wheel. Formerly, axles and wheels were made up of a single unit; separating them made carts much more stable, especially when making turns—and, as noted earlier, greatly increased the mechanical advantage of the wheels themselves.

Transportation entered a new phase in about 2000 B.C., when improvements in technology made possible the development of spoked wheels. By heat-treating wood, it became possible to bend the material slightly, and to attach spokes between the rim and hub of the wheel. When the wood cooled, the tension created a much stronger wheel—capable of carrying heavier loads faster and over greater distances.

In China during the first century B.C., a new type of wheeled vehicle—identified earlier as a

Class II lever—was born in the form of the wheelbarrow, or "wooden ox." The wheelbarrow, whose invention the Chinese attributed to a semi-legendary figure named Ko Yu, was of such value to the imperial army for moving arms and military equipment that China's rulers kept its design secret for centuries.

In Europe, around the same time, chariots were dying out, and it was a long time before the technology of wheeled transport improved. The first real innovation came during the 1500s, with the development of the horsedrawn coach. By 1640, a German family was running a regular stagecoach service, and, in 1667, a new, light, two-wheeled carriage called a cabriolet made its first appearance. Later centuries, of course, saw the development of increasingly more sophisticated varieties of wheeled vehicles powered in turn by human effort (the bicycle), steam (the locomotive), and finally, the internal combustion engine (the automobile.)

But the wheel was never just a machine for transport: long before the first wheeled carts came into existence, potters had been using wheels that rotated in place to fashion perfectly round objects, and in later centuries, wheels gained many new applications. By 500 B.C., farmers in Greece and other parts of the Mediterranean world were using rotary mills powered by donkeys. These could grind grain much faster than a person working with a hand-powered grindstone could hope to do, and in time, the Greeks found a means of powering their mills with a force more useful than donkeys: water.

The first waterwheels, turned by human or animal power, included a series of buckets along the rim that made it possible to raise water from the river below and disperse it to other points. By about 70 B.C., however, Roman engineers recognized that they could use the power of water itself to turn wheels and grind grain. Thus, the waterwheel became one of the first two rotor mechanisms in which an inanimate source (as opposed to the effort of humans or animals) created power to spin a shaft.

In this way, the waterwheel was a prototype for the engine developed many centuries later. Indeed, in the first century A.D., Hero of Alexandria—who discovered the concept of steam power some 1,700 years before anyone took up the idea and put it to use—proposed what has been considered a prototype for the turbine engine. However, for a variety of complex reasons, the ancient world was simply not ready for the technological leap portended by such an invention; and so, in terms of significant progress in the development of machines, Europe was asleep for more than a millennium.

The other significant form of wheel powered by an "inanimate" source was the windmill, first mentioned in 85 B.C. by Antipater of Thessalonica, who commented on a windmill he saw in northern Greece. In this early version of the windmill, the paddle wheel moved on a horizontal plane. However, the windmill did not take hold in Europe during ancient times, and, in fact, its true origins lie further east, and it did not become widespread until much later.

In the seventh century A.D., windmills began to appear in the region of modern Iran and Afghanistan, and the concept spread to the Arab world. Europeans in the Near East during the Crusades (1095-1291) observed the windmill, and brought the idea back to Europe with them. By the twelfth century Europeans had developed the more familiar vertical mill.

Finally, there was a special variety of wheel that made its appearance as early as 500 B.C.: the toothed gearwheel. By 300 B.C., it was in use throughout Egypt, and by about 270 B.C., Ctesibius of Alexandria (fl. c. 270-250 B.C.) had applied the gear in devising a constant-flow water clock called a clepsydra.

Some 2,100 years after Ctesibius, toothed gears became a critical component of industrialization. The most common type is a spur gear, in which the teeth of the wheel are parallel to the axis of rotation. Helical gears, by contrast, have curved teeth in a spiral pattern at an angle to their rotational axes. This means that several teeth of one gearwheel are always in contact with several teeth of the adjacent wheel, thus providing greater torque.

In bevel gears, the teeth are straight, as with a spur gear, but they slope at a 45°-angle relative to their axes so that two gearwheels can fit together at up to 90°-angles to one another without a change in speed. Finally, planetary gears are made such that one or more smaller gearwheels can fit within a larger gearwheel, which has teeth cut on the inside rather than the outside.

Similar in concept to the gearwheel is the V belt drive, which consists of two wheels side by side, joined with a belt. Each of the wheels has grooves cut in it for holding the belt, making this a modification of the pulley, and the grooves provide much greater gripping power for holding the belt in place. One common example of a V belt drive, combined with gearwheels, is a bicycle chain assembly.

PULLEYS. A pulley is essentially a grooved wheel on an axle attached to a frame, which in turn is attached to some form of rigid support such as a ceiling. A rope runs along the grooves of the pulley, and one end is attached to a load while the other is controlled by the operator.

In several instances, it has been noted that a machine may provide increased range of motion or position rather than power. So, this simplest kind of pulley, known as a single or fixed pulley, only offers the advantage of direction rather than improved force. When using Venetian blinds, there is no increase in force; the advantage of the machine is simply that it allows one to move objects upward and downward. Thus, the theoretical mechanical advantage of a fixed pulley is 1 (or almost certainly less under actual conditions, where friction is a factor).

Here it is appropriate to return to Archimedes, whose advancements in the understanding of levers translated to improvements in pulleys. In the case of the lever, it was Archimedes who first recognized that the longer the effort arm, the less effort one had to apply in raising the load. Likewise, with pulleys and related devices—cranes and winches—he explained and improved the way these machines worked.

The first crane device dates to about 1000 B.C., but evidence from pictures suggests that pulleys may have been in use as early as seven thousand years before. Several centuries before Archimedes's time, the Greeks were using compound pulleys that contained several wheels and thus provided the operator with much greater mechanical advantage than a fixed pulley. Archimedes, who was also the first to recognize the relationship between pulleys and levers, created the first fully realized block-and-tackle system using compound pulleys and cranes. In the late modern era, compound pulley systems were used in applications such as elevators and escalators.

A compound pulley consists of two or more wheels, with at least one attached to the support while the other wheel or wheels lift the load. A rope runs from the support pulley down to the load-bearing wheel, wraps around that pulley and comes back up to a fixed attachment on the upper pulley. Whereas the upper pulley is fixed, the load-bearing pulley is free to move, and raises the load as the rope is pulled below.

The simplest kind of compound pulley, with just two wheels, has a mechanical advantage of 2. On a theoretical level, at least, it is possible to calculate the mechanical advantage of a compound pulley with more wheels: the number is equal to the segments of rope between the lower pulleys and the upper, or support pulley. In reality, however, friction, which is high as ropes rub against the pulley wheels, takes its toll. Thus mechanical advantage is never as great as it might be.

A block-and-tackle, like a compound pulley, uses just one rope with a number of pulley wheels. In a block-and-tackle, however, the wheels are arranged along two axles, each of which includes multiple pulley wheels that are free to rotate along the axle. The upper row is attached to the support, and the lower row to the load. The rope connects them all, running from the first pulley in the upper set to the first in the lower set, then to the second in the upper set, and so on. In theory, at least, the mechanical advantage of a block-and-tackle is equal to the number of wheels used, which must be an even number—but again, friction diminishes the theoretical mechanical advantage.

The Inclined Plane

To the contemporary mind, it is difficult enough to think of a lever as a "machine"—but levers at least have more than one part, unlike an inclined plane. The latter, by contrast, is exactly what it seems to be: a ramp. Yet it was just such a ramp structure, as noted earlier, that probably enabled the Egyptians to build the pyramids—a feat of engineering so stunning that even today, some people refuse to believe that the ancient Egyptians could have achieved it on their own.

Surely, as anti-scientific proponents of various fantastic theories often insist, the building of the pyramids could only have been done with machines provided by super-intelligent, extraterrestrial beings. Even in ancient times, the Greek historian Herodotus (c. 484-c. 424 B.C.) speculated that the Egyptians must have used huge cranes that had long since disappeared.

These bizarre guesses concerning the technology for raising the pyramid's giant blocks serve to highlight the brilliance of a gloriously simple machine that, in essence, doubles force. If one needs to move a certain weight to a certain height, there are two options. One can either raise the weight straight upward, expending an enormous amount of effort, even with a pulley system; or one can raise the weight gradually along an inclined plane. The inclined plane is a much wiser choice, because it requires half the effort.

Why half? Imagine an inclined plane sloping evenly upward to the right. The plane exists in a sort of frame that is equal in both length and height to the dimensions of the plane itself. As we can easily visualize, the plane takes up exactly half of the frame, and this is true whether the slope is more than, less than, or equal to 45°. For any plane in which the slope is more than 45°, however, the mechanical advantage will be less than 1, and it is indeed hard to imagine why anyone would use such a plane unless forced to do so by limitations on their horizontal space—for example, when lifting a heavy object from a narrow canyon.

The mechanical advantage of an inclined plane is equal to the ratio between the distance over which input force is applied and the distance of output; or, more simply, the ratio of length to height. If a man is pushing a crate up a ramp 4 ft high and 8 ft long (1.22 m by 2.44 m), 8 ft is the input distance and 4 ft the output distance; hence, the mechanical advantage is 2. If the ramp length were doubled to 16 ft (4.88 m), the mechanical advantage would likewise double to 4, and so on.

The concept of work, in terms of physics, has specific properties that are a subject unto themselves; however, it is important here only to recognize that work is the product of force (that is, effort) multiplied by distance. This means that if one increases the distance, a much smaller quantity of force is needed to achieve the same amount of work.

On an everyday level, it is easy to see this in action. Walking or running up a gentle hill, obviously, is easier than going up a steep hill. Therefore, if one's primary purpose is to conserve effort, it is best to choose the gentler hill. On the other hand, one may wish to minimize distance—or, if moving for the purpose of exercise, to maximize force input to burn calories. In either case, the steeper hill would be the better option.

WEDGES. The type of inclined plane discussed thus far is a ramp, but there are a number of much smaller varieties of inclined plane at work in the everyday world. A knife is an excellent example of one of the most common types, a wedge. Again, the mechanical advantage of a wedge is the ratio of length to height, which, in the knife, would be the depth of the blade compared to its cross-sectional width. Due to the ways in which wedges are used, however, friction plays a much greater role, therefore greatly reducing the theoretical mechanical advantage.

Other types of wedges may be used with a lever, as a form of fulcrum for raising objects. Or, a wedge may be placed under objects to stabilize them, as for instance, when a person puts a folded matchbook under the leg of a restaurant table to stop it from wobbling. Wedges also stop other objects from moving: a triangular piece of wood under a door will keep it from closing, and a more substantial wedge under the front wheels of a car will stop it from rolling forward.

Variations of the wedge are everywhere. Consider all the types of cutting or chipping devices that exist: scissors, chisels, ice picks, axes, splitting wedges (used with a mallet to split a log down the center), saws, plows, electric razors, etc. Then there are devices that use a complex assembly of wedges working together. The part of a key used to open a lock is really just a row of wedges for moving the pins inside the lock to the proper position for opening the door. Similarly, each tooth in a zipper is a tiny wedge that fits tightly with the adjacent teeth.

SCREWS. As the wheel is, without a doubt, the greatest conceptual variation on the lever, so the screw may be identified as a particularly cunning adaptation of an inclined plane. The uses of screws today are many and obvious, but as with wheels and axles, these machines have more applications than are commonly recognized. Not only are there screws for holding things together, but there are screws such as those on vises, clamps, or monkey wrenches for applying force to objects.

A screw is an inclined plane in the shape of a helix, wrapped around an axis or cylinder. In order to determine its mechanical advantage, one must first find the pitch, which is the distance

A screw, like this corkscrew used to open a bottle of wine, is an inclined plane in the shape of a helix, wrapped around an axis or cylinder. *(Ecoscene/Corbis. Reproduced by permission.)*

between adjacent threads. The other variable is lever arm, which with a screwdriver is the radius, or on a wrench, the length from the crescent or clamp to the area of applied force. Obviously, the lever arm is much greater for a wrench, which explains why a wrench is sometimes preferable to a screwdriver, when removing a highly resistant material screw or bolt.

When one rotates a screw of a given pitch, the applied force describes a circle whose area may be calculated as $2\pi L$, where L is the lever arm. This figure, when divided by the pitch, is the same as the ratio between the distance of force input to force output. Either number is equal to the mechanical advantage for a screw. As suggested earlier, that mechanical advantage is usually low, because force input (screwing in the screw) takes place in a much greater range of motion than force output (the screw working its way into the surface). But this is exactly what the screw is designed to do, and what it lacks in mechanical advantage, it more than makes up in its holding power.

As with the lever and pulley, Archimedes did not invent the screw, but he did greatly improve human understanding of it. Specifically, he developed a mathematical formula for a simple spiral, and translated this into the highly practical Archimedes screw, a device for lifting water. The invention consists of a metal pipe in a corkscrew shape, which draws water upward as it revolves. It proved particularly useful for lifting water that had seeped into the lower parts of a ship, and in many countries today, it remains in use as a simple pump for drawing water out of the ground.

Some historians maintain that Archimedes did not invent the screw-type pump, but rather saw an example of it in Egypt. In any case, he clearly developed a practical version of the device, and it soon gained application throughout the ancient world. Archaeologists discovered a screw-driven olive press in the ruins of Pompeii, destroyed by the eruption of Mt. Vesuvius in A.D. 79, and Hero of Alexandria later mentioned the use of a screw-type machine in his *Mechanica.*

Yet, Archimedes is the figure most widely associated with the development of this wondrous device. Hence, in 1837, when the Swedish-American engineer John Ericsson (1803-1899) demonstrated the use of a screw-driven ship's propeller, he did so on a craft he named the Archimedes.

From screws planted in wood to screws that drive ships at sea, the device is everywhere in modern life. Faucets, corkscrews, drills, and meat grinders are obvious examples. Though many types of jacks used for lifting an automobile or a house are levers, others are screw assemblies on which one rotates the handle along a horizontal axis. In fact, the jack is a particularly interesting device. Versions of the jack represent all three types of simple machine: lever, inclined plane, and hydraulic press.

The Hydraulic Press

As noted earlier, the hydraulic press came into existence much, much later than the lever or inclined plane, and its birth can be seen within the context of a larger movement toward the use of water power, including steam. A little more than a quarter-century after Pascal created the theoretical framework for hydraulic power, his countryman Denis Papin (1647-1712) introduced the steam digester, a prototype for the pressure cooker. In 1687, Papin published a work describing a machine in which steam operated a piston—an early model for the steam engine.

Papin's concept, which was on the absolute cutting edge of technological development at that time, utilized not only steam power but also the very hydraulic concept that Pascal had identified a few decades earlier. Indeed, the assembly of pistons and cylinders that forms the central component of the internal-combustion engine reflects this hydraulic rule, discovered by Pascal in 1653. It was then that he formulated what is known as Pascal's principle: that the external pressure applied on a fluid is transmitted uniformly throughout the entire body of that fluid.

Inside a piston and cylinder assembly, one of the most basic varieties of hydraulic press, the pressure is equal to the ratio of force to the horizontal area of pressure. A simple hydraulic press of the variety that might be used to raise a car in an auto shop typically consists of two large cylinders side by side, connected at the bottom by a channel in which valves control flow. When one applies force over a given area of input—that is, by pressing down on one cylinder—this yields a uniform pressure that causes output in the second cylinder.

Once again, mechanical advantage is equal to the ratio of force output to force input, and for a hydraulic press, this can also be measured as the ratio of area output to area input. Just as there is an inverse relationship between lever arm and force in a lever, and between length and height in an inclined plane, so there is such a relationship between horizontal area and force in a hydraulic pump. Consequently, in order to increase force, one should minimize area.

However, there is another factor to consider: height. The mechanical advantage of a hydraulic pump is equal to the vertical distance to which the input force is applied, divided by that of the output force. Hence, the greater the height of the input cylinder compared to the output cylinder, the greater the mechanical advantage. And since these three factors—height, area, and force—work together, it is possible to increase the lifting force and area by minimizing the height.

Consider once again the auto-shop car jack. Typically, the input cylinder will be relatively tall and thin, and the output cylinder short and squat. Because the height of the input cylinder is large, the area of input will be relatively small, as will the force of input. But according to Pascal's principle, whatever the force applied on the input, the pressure will be the same on the output. At the output end, where the car is raised, one needs a large amount of force and a relatively large lifting area. Therefore, height is minimized to increase force and area. If the output area is 10 times the size of the input area, an input force of 1 unit will produce an output force of 10 units—but in order to raise the weight by 1 unit of height, the input piston must move downward by 10 units.

This type of car jack provides a basic model of the hydraulic press in operation, but, in fact, hydraulic technology has many more applications. A hydraulic pump, whether for pumping air into a tire or water from a basement, uses very much the same principle as the hydraulic jack. So too does the hydraulic ram, used in machines ranging from bulldozers to the hydraulic lifts used by firefighters and utility workers to reach great heights.

In a hydraulic ram, however, the characteristics of the input and output cylinders are reversed from those of a car jack. The input cylinder, called the master cylinder, is short and squat, whereas the output cylinder—the slave

KEY TERMS

COMPOUND LEVER: A machine that combines multiple levers to accomplish its task. An example is a piano or manual typewriter.

CLASS I LEVER: A lever in which the fulcrum is between the input force and output force. Examples include a crowbar, a nail puller, and scissors.

CLASS II LEVER: A lever in which the output force is between the input force and the fulcrum. Class II levers, of which wheelbarrows and bottle openers are examples, maximize output force at the expense of range of motion.

CLASS III LEVER: A lever in which the input force is between the output force and the fulcrum. Class III levers, of which a fishing rod is an example, maximize range of motion at the expense of output force.

EFFICIENCY: The ratio of actual mechanical advantage to theoretical mechanical advantage.

FRICTION: The force that resists motion when the surface of one object comes into contact with the surface of another.

FULCRUM: The support point of a lever.

INERTIA: The tendency of an object in motion to remain in motion, and of an object at rest to remain at rest.

INPUT: The effort supplied by the operator of a machine. In a Class I lever such as a crowbar, input would be the energy one expends by pushing down on the bar. Input force is often called applied force, effort force, or simply effort.

LEVER: One of the three basic varieties of machine, a lever consists of a rigid bar supported at one point, known as the fulcrum.

LEVER ARM: On a lever, the distance from the input force or the output force to the fulcrum.

MACHINE: A device that transmits or modifies force or torque for a specific purpose.

MECHANICAL ADVANTAGE: The ratio of force output to force input for a machine.

MOMENT ARM: For an object experiencing torque, moment arm is the distance from the pivot or balance point to the vector on which force is being applied. Moment arm is always perpendicular to the direction of force.

OUTPUT: The results achieved from the operation of a machine. In a Class I lever such as a crowbar, output is the moving of a stone or other heavy load dislodged by the crowbar. Output force is often called the load or resistance force.

TORQUE: In general terms, torque is turning force; in scientific terms, it is the product of moment arm multiplied by force.

cylinder—is tall and thin. The reason for this change is that in objects using a hydraulic ram, height is more important than force output: they are often raising people rather than cars. When the slave cylinder exerts pressure on the stabilizer ram above it (the bucket containing the firefighter, for example), it rises through a much larger range of vertical motion than that of the fluid flowing from the master cylinder.

As noted earlier, the pistons of a car engine are hydraulic pumps—specifically, reciprocating hydraulic pumps, so named because they all work together. In scientific terms, "fluid" can mean either a liquid or a gas such as air; hence, there is an entire subset of hydraulic machines that are pneumatic, or air-powered. Among these are power brakes in a car, pneumatic drills, and even hovercrafts. As with the other two varieties of simple machine, the hydraulic press is in evidence throughout virtually every nook and cranny of daily life. The pump in a toilet tank is a type of hydraulic press, as is the inner chamber of a pen. So too are aerosol cans, fire extinguishers, scuba tanks, water meters, and pressure gauges.

WHERE TO LEARN MORE

Archimedes (Web site). <http://www.mcs.drexel.edu/~crorres/Archimedes/contents.html> (March 9, 2001.)

Bains, Rae. *Simple Machines.* Mahwah, N.J.: Troll Associates, 1985.

Beiser, Arthur. *Physics,* 5th ed. Reading, MA: Addison-Wesley, 1991.

Canizares, Susan. *Simple Machines.* New York: Scholastic, 1999.

Haslam, Andrew and David Glover. *Machines.* Photography by Jon Barnes. Chicago: World Book, 1998.

"*Inventors Toolbox: The Elements of Machines*" (Web site). <http://www.mos.org/sln/Leonardo/InventorsToolbox.html> (March 9, 2001).

Macaulay, David. *The New Way Things Work.* Boston: Houghton Mifflin, 1998.

"*Machines*" (Web site). <http://www.galaxy.net:80/~k12/machines/index.s.html> (March 9, 2001).

"*Motion, Energy, and Simple Machines*" (Web site). <http://www.necc.mass.edu/MRVIS/MR3_13/start.html> (March 9, 2001).

O'Brien, Robert, and the Editors of Life. *Machines.* New York: Time-Life Books, 1964.

"*Simple Machines*" (Web site). <http://sln.fi.edu/qa97/spotlight3/spotlight3.html> (March 9, 2001).

Singer, Charles Joseph, et al., editors. *A History of Technology* (8 vols.) Oxford, England: Clarendon Press, 1954-84.

ENERGY

CONCEPT

As with many concepts in physics, energy—along with the related ideas of work and power—has a meaning much more specific, and in some ways quite different, from its everyday connotation. According to the language of physics, a person who strains without success to pull a rock out of the ground has done no work, whereas a child playing on a playground produces a great deal of work. Energy, which may be defined as the ability of an object to do work, is neither created nor destroyed; it simply changes form, a concept that can be illustrated by the behavior of a bouncing ball.

HOW IT WORKS

In fact, it might actually be more precise to say that energy is the ability of "a thing" or "something" to do work. Not only tangible objects (whether they be organic, mechanical, or electromagnetic) but also non-objects may possess energy. At the subatomic level, a particle with no mass may have energy. The same can be said of a magnetic force field.

One cannot touch a force field; hence, it is not an object—but obviously, it exists. All one has to do to prove its existence is to place a natural magnet, such as an iron nail, within the magnetic field. Assuming the force field is strong enough, the nail will move through space toward it—and thus the force field will have performed work on the nail.

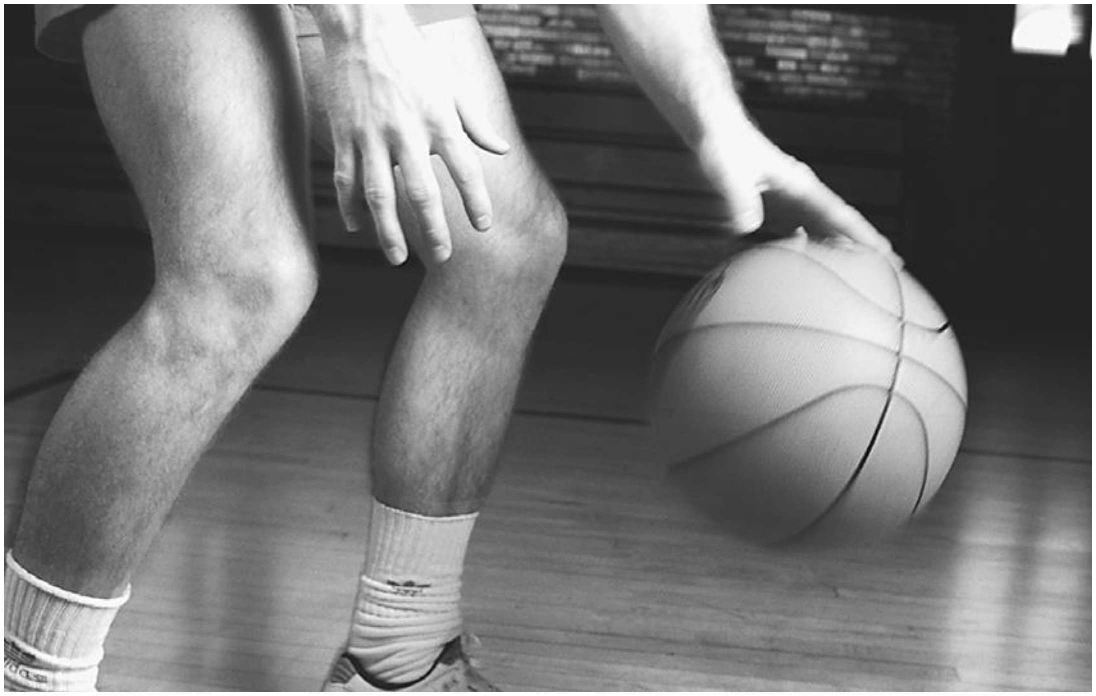
Work: What It Is and Is Not

Work may be defined in general terms as the exertion of force over a given distance. In order for work to be accomplished, there must be a displacement in space—or, in colloquial terms, something has to be moved from point A to point B. As noted earlier, this definition creates results that go against the common-sense definition of "work."

A person straining, and failing, to pull a rock from the ground has performed no work (in terms of physics) because nothing has been moved. On the other hand, a child on a playground performs considerable work: as she runs from the slide to the swing, for instance, she has moved her own weight (a variety of force) across a distance. She is even working when her movement is back-and-forth, as on the swing. This type of movement results in no net displacement, but as long as displacement has occurred at all, work has occurred.

Similarly, when a man completes a full push-up, his body is in the same position—parallel to the floor, arms extended to support him—as he was before he began it; yet he has accomplished work. If, on the other hand, he at the end of his energy, his chest is on the floor, straining but failing, to complete just one more push-up, then he is not working. The fact that he feels as though he has worked may matter in a personal sense, but it does not in terms of physics.

CALCULATING WORK. Work can be defined more specifically as the product of force and distance, where those two vectors are exerted in the same direction. Suppose one were to drag a block of a certain weight across a given distance of floor. The amount of force one exerts parallel to the floor itself, multiplied by the distance, is equal to the amount of work exerted. On the other hand, if one pulls up on the block in a



WHILE THIS PLAYER DRIBBLES HIS BASKETBALL, THE BALL EXPERIENCES A COMPLEX ENERGY TRANSFER AS IT HITS THE FLOOR AND BOUNCES BACK UP. *(Photograph by David Katzenstein/Corbis. Reproduced by permission.)*

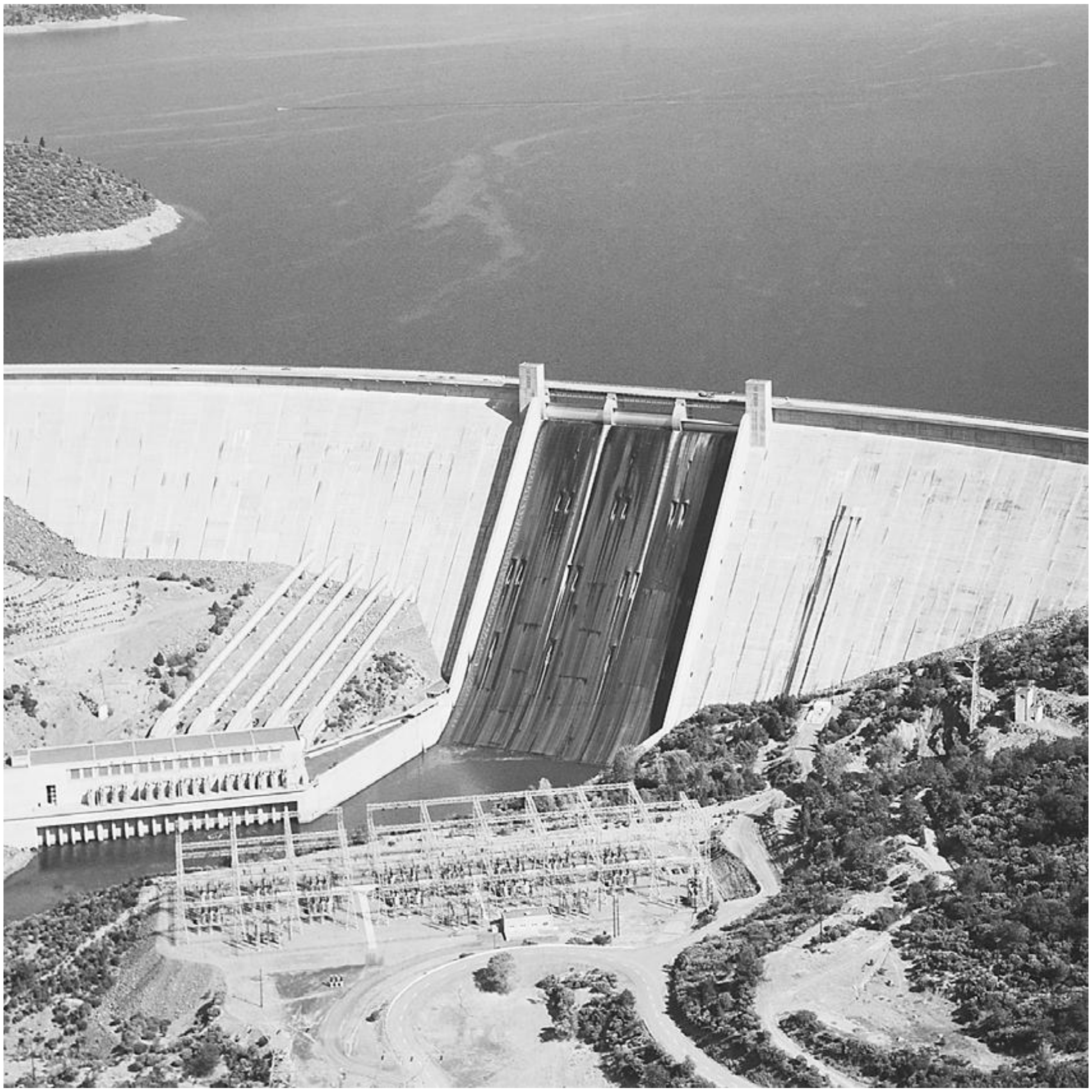
position perpendicular to the floor, that force does not contribute toward the work of dragging the block across the floor, because it is not par allel to distance as defined in this particular situation.

Similarly, if one exerts force on the block at an angle to the floor, only a portion of that force counts toward the net product of work—a portion that must be quantified in terms of trigonometry. The line of force parallel to the floor may be thought of as the base of a triangle, with a line perpendicular to the floor as its second side. Hence there is a 90°-angle, making it a right triangle with a hypotenuse. The hypotenuse is the line of force, which again is at an angle to the floor.

The component of force that counts toward the total work on the block is equal to the total force multiplied by the cosine of the angle. A cosine is the ratio between the leg adjacent to an acute (less than 90°) angle and the hypotenuse. The leg adjacent to the acute angle is, of course, the base of the triangle, which is parallel to the floor itself. Sizes of triangles may vary, but the ratio expressed by a cosine (abbreviated cos) does not. Hence, if one is pulling on the block by a rope that makes a 30°-angle to the floor, then force must be multiplied by cos 30°, which is equal to 0.866.

Note that the cosine is less than 1; hence when multiplied by the total force exerted, it will yield a figure 13.4% smaller than the total force. In fact, the larger the angle, the smaller the cosine; thus for 90°, the value of cos = 0. On the other hand, for an angle of 0°, cos = 1. Thus, if total force is exerted parallel to the floor—that is, at a 0°-angle to it—then the component of force that counts toward total work is equal to the total force. From the standpoint of physics, this would be a highly work-intensive operation.

GRAVITY AND OTHER PECULIARITIES OF WORK. The above discussion relates entirely to work along a horizontal plane. On the vertical plane, by contrast, work is much simpler to calculate due to the presence of a constant downward force, which is, of course, gravity. The force of gravity accelerates objects at a rate of 32 ft (9.8 m)/sec^2. The mass (m) of an object multiplied by the rate of gravitational acceleration (g) yields its weight, and the formula for work done against gravity is equal to weight multiplied by height (h) above some lower reference point: mgh.



Hydroelectric dams, like the Shasta dam in California, superbly illustrate the principle of energy conversion. *(Photograph by Charles E. Rotkin/Corbis. Reproduced by permission.)*

Distance and force are both vectors—that is, quantities possessing both magnitude and direction. Yet work, though it is the product of these two vectors, is a scalar, meaning that only the magnitude of work (and not the direction over which it is exerted) is important. Hence *mgh* can refer either to the upward work one exerts against gravity (that is, by lifting an object to a certain height), or to the downward work that gravity performs on the object when it is dropped. The direction of *h* does not matter, and its value is purely relative, referring to the vertical distance between one point and another.

The fact that gravity can "do work"—and the irrelevance of direction—further illustrates the truth that work, in the sense in which it is applied by physicists, is quite different from "work" as it understood in the day-to-day world. There is a highly personal quality to the everyday meaning of the term, which is completely lacking from its physics definition.

If someone carried a heavy box up five flights of stairs, that person would quite naturally feel justified in saying "I've worked." Certainly he or she would feel that the work expended was far greater than that of someone who had simply allowed the the elevator to carry the box up those five floors. Yet in terms of work done against gravity, the work done on the box by the elevator is exactly the same as that performed by the person carrying it upstairs. The identity of the "worker"—not to mention the sweat expended or not expended—is irrelevant from the standpoint of physics.

Measurement of Work and Power

In the metric system, a newton (N) is the amount of force required to accelerate 1 kg of mass by 1 meter per second squared (m/s^2). Work is measured by the joule (J), equal to 1 newton-meter (N • m). The British unit of force is the pound, and work is measured in foot-pounds, or the work done by a force of 1 lb over a distance of one foot.

Power, the rate at which work is accomplished over time, is the same as work divided by time. It can also be calculated in terms of force multiplied by speed, much like the force-multiplied-by-distance formula for work. However, as with work, the force and speed must be in the same direction. Hence, the formula for power in these terms is $F \cdot \cos\theta \cdot v$, where F=force, v=speed, and $\cos\theta$ is equal to the cosine of the angle θ (the Greek letter theta) between F and the direction of v.

The metric-system measure of power is the watt, named after James Watt (1736-1819), the Scottish inventor who developed the first fully viable steam engine and thus helped inaugurate the Industrial Revolution. A watt is equal to 1 joule per second, but this is such a small unit that it is more typical to speak in terms of kilowatts, or units of 1,000 watts.

Ironically, Watt himself—like most people in the British Isles and America—lived in a world that used the British system, in which the unit of power is the foot-pound per second. The latter, too, is very small, so for measuring the power of his steam engine, Watt suggested a unit based on something quite familiar to the people of his time: the power of a horse. One horsepower (hp) is equal to 550 foot-pounds per second.

SORTING OUT METRIC AND BRITISH UNITS. The British system, of course, is horridly cumbersome compared to the metric system, and thus it long ago fell out of favor with the international scientific community. The British system is the product of loosely developed conventions that emerged over time: for instance, a foot was based on the length of the reigning king's foot, and in time, this became standardized. By contrast, the metric system was created quite deliberately over a matter of just a few years following the French Revolution, which broke out in 1789. The metric system was adopted ten years later.

In this 1957 photograph, Italian opera singer Luigi Infantino tries to break a wine glass by singing a high "C" note. Contrary to popular belief, the note does not have to be a particularly high one to break the glass: rather, the note should be on the same wavelength as the glass's own vibrations. When this occurs, sound energy is transferred directly to the glass, which is shattered by this sudden net intake of energy. *(Hulton-Deutsch Collection/Corbis. Reproduced by permission.)*

During the revolutionary era, French intellectuals believed that every aspect of existence could and should be treated in highly rational, scientific terms. Out of these ideas arose much folly—especially after the supposedly "rational" leaders of the revolution began chopping off people's heads—but one of the more positive outcomes was the metric system. This system, based entirely on the number 10 and its exponents, made it easy to relate one figure to another: for instance, there are 100 centimeters in a meter and 1,000 meters in a kilometer. This is vastly more convenient than converting 12 inches to a foot, and 5,280 feet to a mile.

For this reason, scientists—even those from the Anglo-American world—use the metric system for measuring not only horizontal space, but volume, temperature, pressure, work, power, and so on. Within the scientific community, in fact, the metric system is known as SI, an abbreviation of the French *Système International d'Unités*—that is, "International System of Units."

Americans have shown little interest in adopting the SI system, yet where power is concerned, there is one exception. For measuring the power of a mechanical device, such as an automobile or even a garbage disposal, Americans use the British horsepower. However, for measuring electrical power, the SI kilowatt is used. When an electric utility performs a meter reading on a family's power usage, it measures that usage in terms of electrical "work" performed for the family, and thus bills them by the kilowatt-hour.

Three Types of Energy

KINETIC AND POTENTIAL ENERGY FORMULAE. Earlier, energy was defined as the ability of an object to accomplish work—a definition that by this point has acquired a great deal more meaning. There are three types of energy: kinetic energy, or the energy that something possesses by virtue of its motion; potential energy, the energy it possesses by virtue of its position; and rest energy, the energy it possesses by virtue of its mass.

The formula for kinetic energy is KE = $\frac{1}{2} mv^2$. In other words, for an object of mass m, kinetic energy is equal to half the mass multiplied by the square of its speed v. The actual derivation of this formula is a rather detailed process, involving reference to the second of the three laws of motion formulated by Sir Isaac Newton (1642-1727.) The second law states that $F = ma$, in other words, that force is equal to mass multiplied by acceleration. In order to understand kinetic energy, it is necessary, then, to understand the formula for uniform acceleration. The latter is $v_f^2 = v_0^2 + 2as$, where v_f^2 is the final speed of the object, v_0^2 its initial speed, a acceleration and s distance. By substituting values within these equations, one arrives at the formula of $\frac{1}{2} mv^2$ for kinetic energy.

The above is simply another form of the general formula for work—since energy is, after all, the ability to perform work. In order to produce an amount of kinetic energy equal to $\frac{1}{2} mv^2$ within an object, one must perform an amount of work on it equal to Fs. Hence, kinetic energy also equals Fs, and thus the preceding paragraph simply provides a means for translating that into more specific terms.

The potential energy (PE) formula is much simpler, but it also relates to a work formula given earlier: that of work done against gravity. Potential energy, in this instance, is simply a function of gravity and the distance h above some reference point. Hence, its formula is the same as that for work done against gravity, mgh or wh, where w stands for weight. (Note that this refers to potential energy in a gravitational field; potential energy may also exist in an electromagnetic field, in which case the formula would be different from the one presented here.)

REST ENERGY AND ITS INTRIGUING FORMULA. Finally, there is rest energy, which, though it may not sound very exciting, is in fact the most intriguing—and the most complex—of the three. Ironically, the formula for rest energy is far, far more complex in derivation than that for potential or even kinetic energy, yet it is much more well-known within the popular culture.

Indeed, $E = mc^2$ is perhaps the most famous physics formula in the world—even more so than the much simpler $F = ma$. The formula for rest energy, as many people know, comes from the man whose Theory of Relativity invalidated certain specifics of the Newtonian framework: Albert Einstein (1879-1955). As for what the formula actually means, that will be discussed later.

REAL-LIFE APPLICATIONS

Falling and Bouncing Balls

One of the best—and most frequently used—illustrations of potential and kinetic energy involves standing at the top of a building, holding a baseball over the side. Naturally, this is not an experiment to perform in real life. Due to its relatively small mass, a falling baseball does not have a great amount of kinetic energy, yet in the real world, a variety of other conditions (among them inertia, the tendency of an object to maintain its state of motion) conspire to make a hit on the head with a baseball potentially quite serious.

If dropped from a great enough height, it could be fatal.

When one holds the baseball over the side of the building, potential energy is at a peak, but once the ball is released, potential energy begins to decrease in favor of kinetic energy. The relationship between these, in fact, is inverse: as the value of one decreases, that of the other increases in exact proportion. The ball will only fall to the point where its potential energy becomes 0, the same amount of kinetic energy it possessed before it was dropped. At the same point, kinetic energy will have reached maximum value, and will be equal to the potential energy the ball possessed at the beginning. Thus the sum of kinetic energy and potential energy remains constant, reflecting the conservation of energy, a subject discussed below.

It is relatively easy to understand how the ball acquires kinetic energy in its fall, but potential energy is somewhat more challenging to comprehend. The ball does not really "possess" the potential energy: potential energy resides within an entire system comprised by the ball, the space through which it falls, and the Earth. There is thus no "magic" in the reciprocal relationship between potential and kinetic energy: both are part of a single system, which can be envisioned by means of an analogy.

Imagine that one has a 20-dollar bill, then buys a pack of gum. Now one has, say, \$19.20. The positive value of dollars has decreased by \$0.80, but now one has increased "non-dollars" or "anti-dollars" by the same amount. After buying lunch, one might be down to \$12.00, meaning that "anti-dollars" are now up to \$8.00. The same will continue until the entire \$20.00 has been spent. Obviously, there is nothing magical about this: the 20-dollar bill was a closed system, just like the one that included the ball and the ground. And just as potential energy decreased while kinetic energy increased, so "non-dollars" increased while dollars decreased.

BOUNCING BACK. The example of the baseball illustrates one of the most fundamental laws in the universe, the conservation of energy: within a system isolated from all other outside factors, the total amount of energy remains the same, though transformations of energy from one form to another take place. An interesting example of this comes from the case of another ball and another form of vertical motion.

This time instead of a baseball, the ball should be one that bounces: any ball will do, from a basketball to a tennis ball to a superball. And rather than falling from a great height, this one is dropped through a range of motion ordinary for a human being bouncing a ball. It hits the floor and bounces back—during which time it experiences a complex energy transfer.

As was the case with the baseball dropped from the building, the ball (or more specifically, the system involving the ball and the floor) possesses maximum potential energy prior to being released. Then, in the split-second before its impact on the floor, kinetic energy will be at a maximum while potential energy reaches zero.

So far, this is no different than the baseball scenario discussed earlier. But note what happens when the ball actually hits the floor: it stops for an infinitesimal fraction of a moment. What has happened is that the impact on the floor (which in this example is assumed to be perfectly rigid) has dented the surface of the ball, and this saps the ball's kinetic energy just at the moment when the energy had reached its maximum value. In accordance with the energy conservation law, that energy did not simply disappear: rather, it was transferred to the floor.

Meanwhile, in the wake of its huge energy loss, the ball is motionless. An instant later, however, it reabsorbs kinetic energy from the floor, undents, and rebounds. As it flies upward, its kinetic energy begins to diminish, but potential energy increases with height. Assuming that the person who released it catches it at exactly the same height at which he or she let it go, then potential energy is at the level it was before the ball was dropped.

WHEN A BALL LOSES ITS BOUNCE. The above, of course, takes little account of energy "loss"—that is, the transfer of energy from one body to another. In fact, a part of the ball's kinetic energy will be lost to the floor because friction with the floor will lead to an energy transfer in the form of thermal, or heat, energy. The sound that the ball makes when it bounces also requires a slight energy loss; but friction—a force that resists motion when the surface of one object comes into contact with the surface of another—is the principal culprit where energy transfer is concerned.

Of particular importance is the way the ball responds in that instant when it hits bottom and stops. Hard rubber balls are better suited for this purpose than soft ones, because the harder the rubber, the greater the tendency of the molecules to experience only elastic deformation. What this means is that the spacing between molecules changes, yet their overall position does not.

If, however, the molecules change positions, this causes them to slide against one another, which produces friction and reduces the energy that goes into the bounce. Once the internal friction reaches a certain threshold, the ball is "dead"—that is, unable to bounce. The deader the ball is, the more its kinetic energy turns into heat upon impact with the floor, and the less energy remains for bouncing upward.

Varieties of Energy in Action

The preceding illustration makes several references to the conversion of kinetic energy to thermal energy, but it should be stressed that there are only three fundamental varieties of energy: potential, kinetic, and rest. Though heat is often discussed as a form unto itself, this is done only because the topic of heat or thermal energy is complex: in fact, thermal energy is simply a result of the kinetic energy between molecules.

To draw a parallel, most languages permit the use of only three basic subject-predicate constructions: first person ("I"), second person ("you"), and third person ("he/she/it.") Yet within these are endless varieties such as singular and plural nouns or various temporal orientations of verbs: present ("I go"); present perfect ("I have gone"); simple past ("I went"); past perfect ("I had gone.") There are even "moods," such as the subjunctive or hypothetical, which permit the construction of complex thoughts such as "I would have gone." Yet for all this variety in terms of sentence pattern—actually, a degree of variety much greater than for that of energy types—all subject-predicate constructions can still be identified as first, second, or third person.

One might thus describe thermal energy as a manifestation of energy, rather than as a discrete form. Other such manifestations include electromagnetic (sometimes divided into electrical and magnetic), sound, chemical, and nuclear. The principles governing most of these are similar: for instance, the positive or negative attraction between two electromagnetically charged particles is analogous to the force of gravity.

MECHANICAL ENERGY. One term not listed among manifestations of energy is mechanical energy, which is something different altogether: the sum of potential and kinetic energy. A dropped or bouncing ball was used as a convenient illustration of interactions within a larger system of mechanical energy, but the example could just as easily have been a roller coaster, which, with its ups and downs, quite neatly illustrates the sliding scale of kinetic and potential energy.

Likewise, the relationship of Earth to the Sun is one of potential and kinetic energy transfers: as with the baseball and Earth itself, the planet is pulled by gravitational force toward the larger body. When it is relatively far from the Sun, it possesses a higher degree of potential energy, whereas when closer, its kinetic energy is highest. Potential and kinetic energy can also be illustrated within the realm of electromagnetic, as opposed to gravitational, force: when a nail is some distance from a magnet, its potential energy is high, but as it moves toward the magnet, kinetic energy increases.

ENERGY CONVERSION IN A DAM. A dam provides a beautiful illustration of energy conversion: not only from potential to kinetic, but from energy in which gravity provides the force component to energy based in electromagnetic force. A dam big enough to be used for generating hydroelectric power forms a vast steel-and-concrete curtain that holds back millions of tons of water from a river or other body. The water nearest the top—the "head" of the dam—thus has enormous potential energy.

Hydroelectric power is created by allowing controlled streams of this water to flow downward, gathering kinetic energy that is then transferred to powering turbines. Dams in popular vacation spots often release a certain amount of water for recreational purposes during the day. This makes it possible for rafters, kayakers, and others downstream to enjoy a relatively fast-flowing river. (Or, to put it another way, a stream with high kinetic energy.) As the day goes on, however, the sluice-gates are closed once again to build up the "head." Thus when night comes, and energy demand is relatively high as people retreat to their homes, vacation cabins, and hotels, the dam is ready to provide the power they need.

OTHER MANIFESTATIONS OF ENERGY. Thermal and electromagnetic energy are much more readily recognizable manifestations of energy, yet sound and chemical energy are two forms that play a significant part as well. Sound, which is essentially nothing more than the series of pressure fluctuations within a medium such as air, possesses enormous energy: consider the example of a singer hitting a certain note and shattering a glass.

Contrary to popular belief, the note does not have to be particularly high: rather, the note should be on the same wavelength as the glass's own vibrations. When this occurs, sound energy is transferred directly to the glass, which is shattered by this sudden net intake of energy. Sound waves can be much more destructive than that: not only can the sound of very loud music cause permanent damage to the ear drums, but also, sound waves of certain frequencies and decibel levels can actually drill through steel. Indeed, sound is not just a by-product of an explosion; it is part of the destructive force.

As for chemical energy, it is associated with the pull that binds together atoms within larger molecular structures. The formation of water molecules, for instance, depends on the chemical bond between hydrogen and oxygen atoms. The combustion of materials is another example of chemical energy in action.

With both chemical and sound energy, however, it is easy to show how these simply reflect the larger structure of potential and kinetic energy discussed earlier. Hence sound, for instance, is potential energy when it emerges from a source, and becomes kinetic energy as it moves toward a receiver (for example, a human ear). Furthermore, the molecules in a combustible material contain enormous chemical potential energy, which becomes kinetic energy when released in a fire.

Rest Energy and Its Nuclear Manifestation

Nuclear energy is similar to chemical energy, though in this instance, it is based on the binding of particles within an atom and its nucleus. But it is also different from all other kinds of energy, because its force component is neither gravitational nor electromagnetic, but based on one of two other known varieties of force: strong nuclear and weak nuclear. Furthermore, nuclear energy—to a much greater extent than thermal or chemical energy—involves not only kinetic and potential energy, but also the mysterious, extraordinarily powerful, form known as rest energy.

Throughout this discussion, there has been little mention of rest energy; yet it is ever-present. Kinetic and potential energy rise and fall with respect to one another; but rest energy changes little. In the baseball illustration, for instance, the ball had the same rest energy at the top of the building as it did in flight—the same rest energy, in fact, that it had when sitting on the ground. And its rest energy is enormous.

NUCLEAR WARFARE. This brings back the subject of the rest energy formula: $E = mc^2$, famous because it made possible the creation of the atomic bomb. The latter, which fortunately has been detonated in warfare only twice in history, brought a swift end to World War II when the United States unleashed it against Japan in August 1945. From the beginning, it was clear that the atom bomb possessed staggering power, and that it would forever change the way nations conducted their affairs in war and peace.

Yet the atom bomb involved only nuclear fission, or the splitting of an atom, whereas the hydrogen bomb that appeared just a few years after the end of World War II used an even more powerful process, the nuclear fusion of atoms. Hence, the hydrogen bomb upped the ante to a much greater extent, and soon the two nuclear superpowers—the United States and the Soviet Union—possessed the power to destroy most of the life on Earth.

The next four decades were marked by a superpower struggle to control "the bomb" as it came to be known—meaning any and all nuclear weapons. Initially, the United States controlled all atomic secrets through its heavily guarded Manhattan Project, which created the bombs used against Japan. Soon, however, spies such as Julius and Ethel Rosenberg provided the Soviets with U.S. nuclear secrets, ensuring that the dictatorship of Josef Stalin would possess nuclear capabilities as well. (The Rosenbergs were executed for treason, and their alleged innocence became a celebrated cause among artists and intellectuals; however, Soviet documents released since the collapse of the Soviet empire make it clear that they were guilty as charged.)

KEY TERMS

CONSERVATION OF ENERGY: A law of physics which holds that within a system isolated from all other outside factors, the total amount of energy re-mains the same, though transformations of energy from one form to another take place.

COSINE: For an acute (less than 90°) in a right triangle, the cosine (abbreviated cos) is the ratio between the adjacent leg and the hypotenuse. Regardless of the size of the triangle, this figure is a constant for any particular angle.

ENERGY: The ability of an object (or in some cases a non-object, such as a magnetic force field) to accomplish work.

FRICTION: The force that resists motion when the surface of one object comes into contact with the surface of another.

HORSEPOWER: The British unit of power, equal to 550 foot-pounds per second.

HYPOTENUSE: In a right triangle, the side opposite the right angle.

JOULE: The SI measure of work. One joule (1 J) is equal to the work required to accelerate 1 kilogram of mass by 1 meter per second squared (1 m/s^2) over a distance of 1 meter. Due to the small size of the joule, however, it is often replaced by the kilowatt-hour, equal to 3.6 million (3.6 • 106) J.

KINETIC ENERGY: The energy that an object possesses by virtue of its motion.

MATTER: Physical substance that occupies space, has mass, is composed of atoms (or in the case of subatomic particles, is part of an atom), and is convertible into energy.

MECHANICAL ENERGY: The sum of potential energy and kinetic energy within a system.

POTENTIAL ENERGY: The energy that an object possesses by virtue of its position.

POWER: The rate at which work is accomplished over time, a figure rendered mathematically as work divided by time.

Both nations began building up missile arsenals. It was not, however, just a matter of the United States and the Soviet Union. By the 1970s, there were at least three other nations in the "nuclear club": Britain, France, and China. There were also other countries on the verge of developing nuclear bombs, among them India and Israel. Furthermore, there was a great threat that a terrorist leader such as Libya's Muammar al-Qaddafi would acquire nuclear weapons and do the unthinkable: actually use them.

Though other nations acquired nuclear weapons, however, the scale of the two superpower arsenals dwarfed all others. And at the heart of the U.S.-Soviet nuclear competition was a sort of high-stakes chess game—to use a metaphor mentioned frequently during the 1970s. Soviet leaders and their American counterparts both recognized that it would be the end of the world if either unleashed their nuclear weapons; yet each was determined to be able to meet the other's ever-escalating nuclear threat.

United States President Ronald Reagan earned harsh criticism at home for his nuclear buildup and his hard line in negotiations with Soviet President Mikhail Gorbachev; but as a result of this one-upmanship, he put the Soviets into a position where they could no longer compete. As they put more and more money into nuclear weapons, they found themselves less and

KEY TERMS CONTINUED

The SI unit of power is the watt, while the British unit is the foot-pound per second. The latter, because it is small, is usually reckoned in terms of horsepower.

REST ENERGY: The energy an object possesses by virtue of its mass.

RIGHT TRIANGLE: A triangle that includes a right (90°) angle. The other two angles are, by definition, acute or less than 90°.

SCALAR: A quantity that possesses only magnitude, with no specific direction.

SI: An abbreviation of the French *Système International d'Unités,* which means "International System of Units." This is the term within the scientific community for the entire metric system, as applied to a wide variety of quantities ranging from length, weight and volume to work and power, as well as electromagnetic units.

SYSTEM: In discussions of energy, the term "system" refers to a closed set of interactions free from interference by outside factors. An example is the baseball dropped from a height to illustrate potential energy and kinetic energy the ball, the space through which it falls, and the ground below together form a system.

VECTOR: A quantity that possesses both magnitude and direction.

WATT: The metric unit of power, equal to 1 joule per second. Because this is such a small unit, scientists and engineers typically speak in terms of kilowatts, or units of 1,000 watts.

WORK: The exertion of force over a given distance. Work is the product of force and distance, where force and distance are exerted in the same direction. Hence the actual formula for work is $F \cdot \cos \theta \cdot s$, where F = force, s = distance, and $\cos \theta$ is equal to the cosine of the angle θ (the Greek letter theta) between F and s. In the metric or SI system, work is measured by the joule (J), and in the British system by the foot-pound.

less able to uphold their already weak economic system. This was precisely Reagan's purpose in using American economic might to outspend the Soviets—or, in the case of the proposed multi-trillion-dollar Strategic Defense Initiative (SDI or "Star Wars")—threatening to outspend them. The Soviets expended much of their economic energy in competing with U.S. military strength, and this (along with a number of other complex factors), spelled the beginning of the end of the Communist empire.

$E = mc^2$. The purpose of the preceding historical brief is to illustrate the epoch-making significance of a single scientific formula: $E = mc^2$. It ended World War II and ensured that no war like it would ever happen again—but brought on the specter of global annihilation. It created a superpower struggle—yet it also ultimately helped bring about the end of Soviet totalitarianism, thus opening the way for a greater level of peace and economic and cultural exchange than the world has ever known. Yet nuclear arsenals still remain, and the nuclear threat is far from over.

So just what is this literally earth-shattering formula? E stands for rest energy, m for mass, and c for the speed of light, which is 186,000 mi (297,600 km) per second. Squared, this yields an almost unbelievably staggering number.

Hence, even an object of insignificant mass possesses an incredible amount of rest energy. The baseball, for instance, weighs only about 0.333 lb, which—on Earth, at least—converts to 0.15 kg. (The latter is a unit of mass, as opposed to weight.) Yet when factored into the rest energy equation, it yields about 3.75 billion kilowatt-hours—enough to provide an American home with enough electrical power to last it more than 156,000 years!

How can a mere baseball possess such energy? It is not the baseball in and of itself, but its mass; thus every object with mass of any kind possesses rest energy. Often, mass energy can be released in very small quantities through purely thermal or chemical processes: hence, when a fire burns, an almost infinitesimal portion of the matter that went into making the fire is converted into energy. If a stick of dynamite that weighed 2.2 lb (1 kg) exploded, the portion of it that "disappeared" would be equal to 6 parts out of 100 billion; yet that portion would cause a blast of considerable proportions.

As noted much earlier, the derivation of Einstein's formula—and, more to the point, how he came to recognize the fundamental principles involved—is far beyond the scope of this essay. What is important is the fact, hypothesized by Einstein and confirmed in subsequent experiments, that matter is convertible to energy, a fact that becomes apparent when matter is accelerated to speeds close to that of light.

Physicists do not possess a means for propelling a baseball to a speed near that of light—or of controlling its behavior and capturing its energy. Instead, atomic energy—whether of the wartime or peacetime varieties (that is, in power plants)—involves the acceleration of mere atomic particles. Nor is any atom as good as another. Typically physicists use uranium and other extremely rare minerals, and often, they further process these minerals in highly specialized ways. It is the rarity and expense of those minerals, incidentally—not the difficulty of actually putting atomic principles to work—that has kept smaller nations from developing their own nuclear arsenals.

WHERE TO LEARN MORE

Beiser, Arthur. *Physics,* 5th ed. Reading, MA: Addison-Wesley, 1991.

Berger, Melvin. *Sound, Heat and Light: Energy at Work.* Illustrated by Anna DiVito. New York: Scholastic, 1992.

Gardner, Robert. *Energy Projects for Young Scientists.* New York: F. Watts, 1987.

"Kinetic and Potential Energy" Thinkquest (Web site). <http://library.thinkquest.org/2745/data/ke.htm> (March 12, 2001).

Snedden, Robert. *Energy.* Des Plaines, IL: Heinemann, Library, 1999.

Suplee, Curt. *Everyday Science Explained.* Washington, D.C.: National Geographic Society, 1996.

"Work and Energy" (Web site). <http://www.glenbrook.k12.il.us/gbssci/phys/Class/energy/energtoc.html> (March 12, 2001).

World of Coasters (Web site). <http://www.worldofcoasters.com> (March 12, 2001).

Zubrowski, Bernie. *Raceways: Having Fun with Balls and Tracks.* Illustrated by Roy Doty. New York: Morrow, 1985.

THERMODYNAMICS

GAS LAWS

CONCEPT

Gases respond more dramatically to temperature and pressure than do the other three basic types of matter (liquids, solids and plasma). For gases, temperature and pressure are closely related to volume, and this allows us to predict their behavior under certain conditions. These predictions can explain mundane occurrences, such as the fact that an open can of soda will soon lose its fizz, but they also apply to more dramatic, life-and-death situations.

HOW IT WORKS

Ordinary air pressure at sea level is equal to 14.7 pounds per square inch, a quantity referred to as an atmosphere (atm). Because a pound is a unit of force and a kilogram a unit of mass, the metric equivalent is more complex in derivation. A newton (N), or 0.2248 pounds, is the metric unit of force, and a pascal (Pa)—1 newton per square meter—the unit of pressure. Hence, an atmosphere, expressed in metric terms, is 1.013×10^5 Pa.

Gases vs. Solids and Liquids: A Strikingly Different Response

Regardless of the units you use, however, gases respond to changes in pressure and temperature in a remarkably different way than do solids or liquids. Using a small water sample, say, 0.2642 gal (1 l), an increase in pressure from 1-2 atm will decrease the volume of the water by less than 0.01%. A temperature increase from 32° to 212°F (0 to 100°C) will increase its volume by only 2% The response of a solid to these changes is even less dramatic; however, the reaction of air (a combination of oxygen, nitrogen, and other gases) to changes in pressure and temperature is radically different.

For air, an equivalent temperature increase would result in a volume increase of 37%, and an equivalent pressure increase will decrease the volume by a whopping 50%. Air and other gases also have a boiling point below room temperature, whereas the boiling point for water is higher than room temperature and that of solids is much higher. The reason for this striking difference in response can be explained by comparing all three forms of matter in terms of their overall structure, and in terms of their molecular behavior. (Plasma, a gas-like state found, for instance, in stars and comets' tails, does not exist on Earth, and therefore it will not be included in the comparisons that follow.)

Molecular Structure Determines Reaction

Solids possess a definite volume and a definite shape, and are relatively noncompressible: for instance, if you apply extreme pressure to a steel plate, it will bend, but not much. Liquids have a definite volume, but no definite shape, and tend to be noncompressible. Gases, on the other hand, possess no definite volume or shape, and are compressible.

At the molecular level, particles of solids tend to be definite in their arrangement and close in proximity—indeed, part of what makes a solid "solid," in the everyday meaning of that term, is the fact that its constituent parts are basically immovable. Liquid molecules, too, are close in proximity, though random in arrangement. Gas molecules, too, are random in arrangement, but tend to be more widely spaced than liquid molecules. Solid particles are slow moving, and have a strong attraction to one another, whereas gas particles are fast-moving, and have little or no attraction. (Liquids are moderate in both regards.)

Given these interesting characteristics of gases, it follows that a unique set of parameters—collectively known as the "gas laws"—are needed to describe and predict their behavior. Most of the gas laws were derived during the eighteenth and nineteenth centuries by scientists whose work is commemorated by the association of their names with the laws they discovered. These men include the English chemists Robert Boyle (1627-1691), John Dalton (1766-1844), and William Henry (1774-1836); the French physicists and chemists J. A. C. Charles (1746-1823) and Joseph Gay-Lussac (1778-1850), and the Italian physicist Amedeo Avogadro (1776-1856).

Boyle's, Charles's, and Gay-Lussac's Laws

Boyle's law holds that in isothermal conditions (that is, a situation in which temperature is kept constant), an inverse relationship exists between the volume and pressure of a gas. (An inverse relationship is a situation involving two variables, in which one of the two increases in direct proportion to the decrease in the other.) In this case, the greater the pressure, the less the volume and vice versa. Therefore the product of the volume multiplied by the pressure remains constant in all circumstances.

Charles's law also yields a constant, but in this case the temperature and volume are allowed to vary under isobarometric conditions—that is, a situation in which the pressure remains the same. As gas heats up, its volume increases, and when it cools down, its volume reduces accordingly. Hence, Charles established that the ratio of temperature to volume is constant.

By now a pattern should be emerging: both of the aforementioned laws treat one parameter (temperature in Boyle's, pressure in Charles's) as unvarying, while two other factors are treated as variables. Both in turn yield relationships between the two variables: in Boyle's law, pressure and volume are inversely related, whereas in Charles's law, temperature and volume are directly related.

In Gay-Lussac's law, a third parameter, volume, is treated as a constant, and the result is a constant ratio between the variables of pressure and temperature. According to Gay-Lussac's law, the pressure of a gas is directly related to its absolute temperature.

Absolute temperature refers to the Kelvin scale, established by William Thomson, Lord Kelvin (1824-1907). Drawing on Charles's discovery that gas at 0°C (32°F) regularly contracted by about 1/273 of its volume for every Celsius degree drop in temperature, Thomson derived the value of absolute zero (-273.15°C or -459.67°F). Using the Kelvin scale of absolute temperature, Gay-Lussac found that at lower temperatures, the pressure of a gas is lower, while at higher temperatures its pressure is higher. Thus, the ratio of pressure to temperature is a constant.

Avogadro's Law

Gay-Lussac also discovered that the ratio in which gases combine to form compounds can be expressed in whole numbers: for instance, water is composed of one part oxygen and two parts hydrogen. In the language of modern science, this would be expressed as a relationship between molecules and atoms: one molecule of water contains one oxygen atom and two hydrogen atoms.

In the early nineteenth century, however, scientists had yet to recognize a meaningful distinction between atoms and molecules. Avogadro was the first to achieve an understanding of the difference. Intrigued by the whole-number relationship discovered by Gay-Lussac, Avogadro reasoned that one liter of any gas must contain the same number of particles as a liter of another gas. He further maintained that gas consists of particles—which he called molecules—that in turn consist of one or more smaller particles.

In order to discuss the behavior of molecules, it was necessary to establish a large quantity as a basic unit, since molecules themselves are very small. For this purpose, Avogadro established the mole, a unit equal to 6.022137×10^{23} (more than 600 billion trillion) molecules. The term "mole" can be used in the same way we use the word "dozen." Just as "a dozen" can refer to twelve cakes or twelve chickens, so "mole" always describes the same number of molecules.

Just as one liter of water, or one liter of mercury, has a certain mass, a mole of any given substance has its own particular mass, expressed in grams. The mass of one mole of iron, for instance, will always be greater than that of one mole of oxygen. The ratio between them is exactly the same as the ratio of the mass of one iron atom to one oxygen atom. Thus the mole makes if possible to compare the mass of one element or one compound to that of another.

Avogadro's law describes the connection between gas volume and number of moles. According to Avogadro's law, if the volume of gas is increased under isothermal and isobarometric conditions, the number of moles also increases. The ratio between volume and number of moles is therefore a constant.

The Ideal Gas Law

Once again, it is easy to see how Avogadro's law can be related to the laws discussed earlier, since they each involve two or more of the four parameters: temperature, pressure, volume, and quantity of molecules (that is, number of moles). In fact, all the laws so far described are brought together in what is known as the ideal gas law, sometimes called the combined gas law.

The ideal gas law can be stated as a formula, $pV = nRT$, where p stands for pressure, V for volume, n for number of moles, and T for temperature. R is known as the universal gas constant, a figure equal to 0.0821 atm • liter/mole • K. (Like most terms in physics, this one is best expressed in metric rather than English units.)

Given the equation $pV = nRT$ and the fact that R is a constant, it is possible to find the value of any one variable—pressure, volume, number of moles, or temperature—as long as one knows the value of the other three. The ideal gas law also makes it possible to discern certain relations: thus if a gas is in a relatively cool state, the product of its pressure and volume is proportionately low; and if heated, its pressure and volume product increases correspondingly. Thus

$$\frac{p_1 V_1}{T_1} = \frac{p_2 V_2}{T_2},$$

where p_1V_1 is the product of its initial pressure and its initial volume, T_1 its initial temperature, p_2V_2 the product of its final volume and final pressure, and T_2 its final temperature.

A fire extinguisher contains a high-pressure mixture of water and carbon dioxide that rushes out of the siphon tube, which is opened when the release valve is depressed. *(Photograph by Craig Lovell/Corbis. Reproduced by permission.)*

Five Postulates Regarding the Behavior of Gases

Five postulates can be applied to gases. These more or less restate the terms of the earlier discussion, in which gases were compared to solids and liquids; however, now those comparisons can be seen in light of the gas laws.

First, the size of gas molecules is minuscule in comparison to the distance between them, making gas highly compressible. In other words, there is a relatively high proportion of empty space between gas molecules.

Second, there is virtually no force attracting gas molecules to one another.

Third, though gas molecules move randomly, frequently colliding with one another, their net effect is to create uniform pressure.

A hot-air balloon floats because the air inside it is not as dense than the air outside. The way in which the density of the air in the balloon is reduced reflects the gas laws. *(Duomo/Corbis. Reproduced by permission.)*

Fourth, the elastic nature of the collisions results in no net loss of kinetic energy, the energy that an object possesses by virtue of its motion. If a stone is dropped from a height, it rapidly builds kinetic energy, but upon hitting a nonelastic surface such as pavement, most of that kinetic energy is transferred to the pavement. In the case of two gas molecules colliding, however, they simply bounce off one another, only to collide with other molecules and so on, with no kinetic energy lost.

Fifth, the kinetic energy of all gas molecules is directly proportional to the absolute temperature of the gas.

Laws of Partial Pressure

Two gas laws describe partial pressure. Dalton's law of partial pressure states that the total pressure of a gas is equal to the sum of its par tial pressures—that is, the pressure exerted by each component of the gas mixture. As noted earlier, air is composed mostly of nitrogen and oxygen. Along with these are small components carbon dioxide and gases collectively known as the rare or noble gases: argon, helium, krypton, neon, radon, and xenon. Hence, the total pressure of a given quantity of air is equal to the sum of the pressures exerted by each of these gases.

Henry's law states that the amount of gas dissolved in a liquid is directly proportional to the partial pressure of the gas above the surface of the solution. This applies only to gases such as oxygen and hydrogen that do not react chemically to liquids. On the other hand, hydrochloric acid will ionize when introduced to water: one or more of its electrons will be removed, and its atoms will convert to ions, which are either positive or negative in charge.

REAL-LIFE APPLICATIONS

Pressure Changes

OPENING A SODA CAN. Inside a can or bottle of carbonated soda is carbon dioxide gas (CO_2), most of which is dissolved in the drink itself. But some of it is in the space (sometimes referred to as "head space") that makes up the difference between the volume of the soft drink and the volume of the container.

At the bottling plant, the soda manufacturer adds high-pressure carbon dioxide to the head space in order to ensure that more CO_2 will be absorbed into the soda itself. This is in accordance with Henry's law: the amount of gas (in this case CO_2) dissolved in the liquid (soda) is directly proportional to the partial pressure of the gas above the surface of the solution—that is, the CO_2 in the head space. The higher the pressure of the CO_2 in the head space, the greater the amount of CO_2 in the drink itself; and the greater the CO_2 in the drink, the greater the "fizz" of the soda.

Once the container is opened, the pressure in the head space drops dramatically. Once again, Henry's law indicates that this drop in pressure will be reflected by a corresponding drop in the amount of CO_2 dissolved in the soda. Over a period of time, the soda will release that gas, and will eventually go "flat."

FIRE EXTINGUISHERS. A fire extinguisher consists of a long cylinder with an operating lever at the top. Inside the cylinder is a tube of carbon dioxide surrounded by a quantity of water, which creates pressure around the CO_2 tube. A siphon tube runs vertically along the length of the extinguisher, with one opening near the bottom of the water. The other end opens in a chamber containing a spring mechanism attached to a release valve in the CO_2 tube.

The water and the CO_2 do not fill the entire cylinder: as with the soda can, there is "head space," an area filled with air. When the operating lever is depressed, it activates the spring mechanism, which pierces the release valve at the top of the CO_2 tube. When the valve opens, the CO_2 spills out in the "head space," exerting pressure on the water. This high-pressure mixture of water and carbon dioxide goes rushing out of the siphon tube, which was opened when the release valve was depressed. All of this happens, of course, in a fraction of a second—plenty of time to put out the fire.

AEROSOL CANS. Aerosol cans are similar in structure to fire extinguishers, though with one important difference. As with the fire extinguisher, an aerosol can includes a nozzle that depresses a spring mechanism, which in turn allows fluid to escape through a tube. But instead of a gas cartridge surrounded by water, most of the can's interior is made up of the product (for instance, deodorant), mixed with a liquid propellant.

The "head space" of the aerosol can is filled with highly pressurized propellant in gas form, and in accordance with Henry's law, a corresponding proportion of this propellant is dissolved in the product itself. When the nozzle is depressed, the pressure of the propellant forces the product out through the nozzle.

A propellant, as its name implies, propels the product itself through the spray nozzle when the latter is depressed. In the past, chlorofluorocarbons (CFCs)—manufactured compounds containing carbon, chlorine, and fluorine atoms—were the most widely used form of propellant. Concerns over the harmful effects of CFCs on the environment, however, has led to the development of alternative propellants, most notably hydrochlorofluorocarbons (HCFCs), CFC-like compounds that also contain hydrogen atoms.

When the Temperature Changes

A number of interesting things, some of them unfortunate and some potentially lethal, occur when gases experience a change in temperature. In these instances, it is possible to see the gas laws—particularly Boyle's and Charles's—at work.

There are a number of examples of the disastrous effects that result from an increase in the temperature of a product containing combustible gases, as with natural gas and petroleum-based products. In addition, the pressure on the gases in aerosol cans makes the cans highly explosive—so much so that discarded cans at a city dump may explode on a hot summer day. Yet there are other instances when heating a gas can produce positive effects.

A hot-air balloon, for instance, floats because the air inside it is not as dense than the air outside. By itself, this fact does not depend on any of the gas laws, but rather reflects the concept of buoyancy. However, the way in which the density of the air in the balloon is reduced does indeed reflect the gas laws.

According to Charles's law, heating a gas will increase its volume. Also, as noted in the first and second propositions regarding the behavior of gases, gas molecules are highly nonattractive to one another, and therefore, there is a great deal of space between them. The increase in volume makes that space even greater, leading to a significant difference in density between the air in the balloon and the air outside. As a result, the balloon floats, or becomes buoyant.

Although heating a gas can be beneficial, cooling a gas is not always a wise idea. If someone were to put a bag of potato chips into a freezer, thinking this would preserve their flavor, he would be in for a disappointment. Much of what maintains the flavor of the chips is the pressurization of the bag, which ensures a consistent internal environment in which preservative chemicals, added during the manufacture of the chips, can keep them fresh. Placing the bag in the freezer causes a reduction in pressure, as per Gay-Lussac's law, and the bag ends up a limp version of its earlier self.

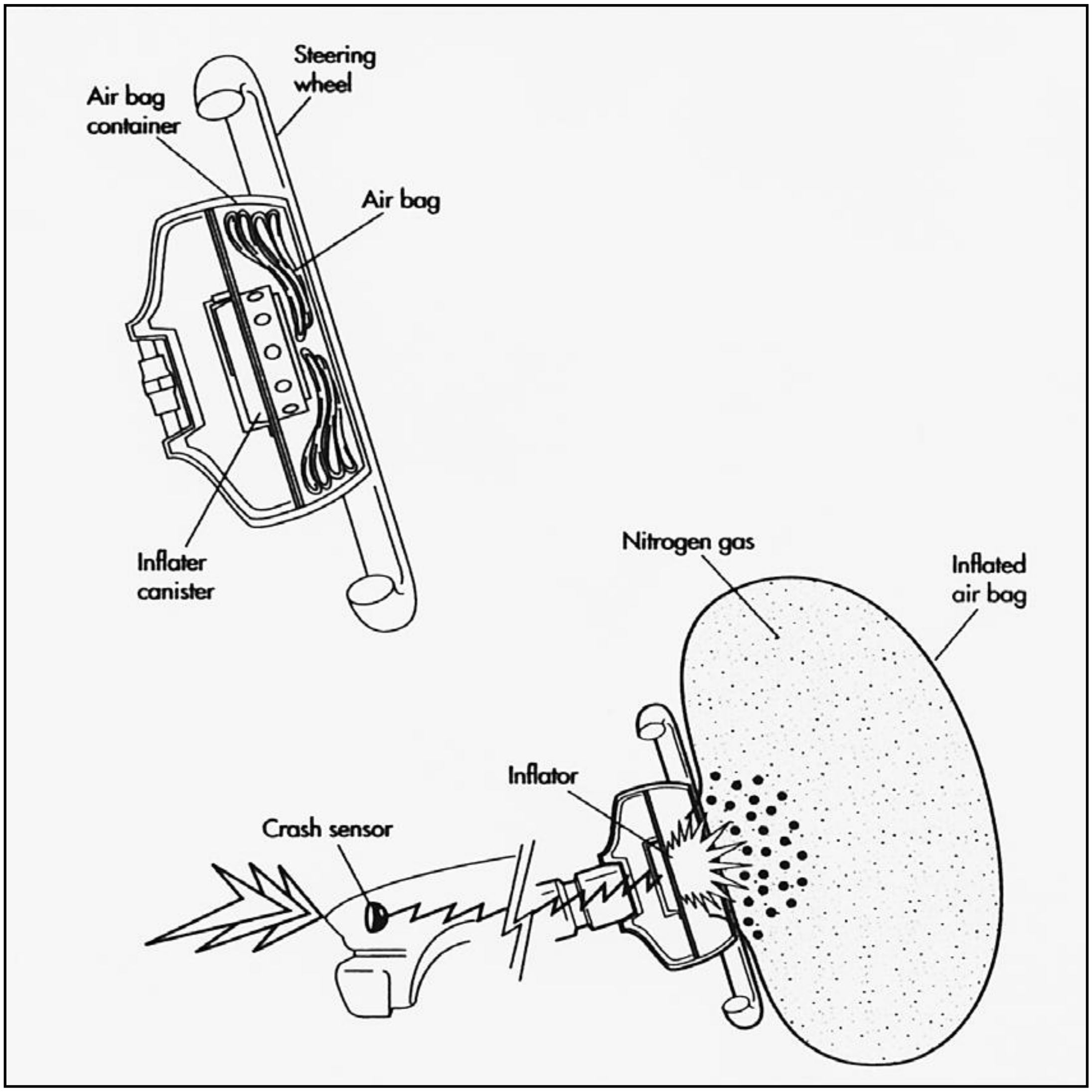
Propane tanks and tires offer an example of the pitfalls that may occur by either allowing a gas to heat up or cool down by too much. Because most propane tanks are made according to strict regulations, they are generally safe, but it is not entirely inconceivable that an extremely hot summer day could cause a defective tank to burst. Certainly the laws of physics are there: an increase in temperature leads to an increase in pressure, in accordance with Gay-Lussac's law, and could lead to an explosion.

Because of the connection between heat and pressure, propane trucks on the highways during the summer are subjected to weight tests to ensure that they are not carrying too much of the gas. On the other hand, a drastic reduction in temperature could result in a loss in gas pressure. If a propane tank from Florida were transported by truck during the winter to northern Canada, the pressure would be dramatically reduced by the time it reached its destination.

Gas Reactions That Move and Stop a Car

In operating a car, we experience two examples of gas laws in operation. One of these, common to everyone, is that which makes the car run: the combustion of gases in the engine. The other is, fortunately, a less frequent phenomenon—but it can and does save lives. This is the operation of an air bag, which, though it is partly related to laws of motion, depends also on the behaviors explained in Charles's law.

With regard to the engine, when the driver pushes down on the accelerator, this activates a throttle valve that sprays droplets of gasoline mixed with air into the engine. (Older vehicles used a carburetor to mix the gasoline and air, but most modern cars use fuel-injection, which sprays the air-gas combination without requiring an intermediate step.) The mixture goes into the



In case of a car collision, a sensor triggers the air bag to inflate rapidly with nitrogen gas. Before your body reaches the bag, however, it has already begun deflating. *(Illustration by Hans & Cassidy. The Gale Group.)*

cylinder, where the piston moves up, compressing the gas and air.

While the mixture is still compressed (high pressure, high density), an electric spark plug produces a flash that ignites it. The heat from this controlled explosion increases the volume of air, which forces the piston down into the cylinder. This opens an outlet valve, causing the piston to rise and release exhaust gases.

As the piston moves back down again, an inlet valve opens, bringing another burst of gasoline-air mixture into the chamber. The piston, whose downward stroke closed the inlet valve, now shoots back up, compressing the gas and air to repeat the cycle. The reactions of the gasoline and air are what move the piston, which turns a crankshaft that causes the wheels to rotate.

So much for moving—what about stopping? Most modern cars are equipped with an airbag, which reacts to sudden impact by inflating. This protects the driver and front-seat passenger, who, even if they are wearing seatbelts, may otherwise be thrown against the steering wheel or dashboard..

But an airbag is much more complicated than it seems. In order for it to save lives, it must deploy within 40 milliseconds (0.04 seconds). Not only that, but it has to begin deflating before the body hits it. An airbag does not inflate if a car simply goes over a bump; it only operates in sit-

KEY TERMS

ABSOLUTE TEMPERATURE: Temperature in relation to absolute zero (-273.15°C or -459.67°F). Its unit is the Kelvin (K), named after William Thomson, Lord Kelvin (1824-1907), who created the scale. The Kelvin and Celsius scales are directly related; hence, Celsius temperatures can be converted to Kelvins (for which neither the word or symbol for "degree" are used) by adding 273.15.

AVOGADRO'S LAW: A statement, derived by the Italian physicist Amedeo Avogadro (1776-1856), which holds that as the volume of gas increases under isothermal and isobarometric conditions, the number of molecules (expressed in terms of mole number), increases as well. Thus the ratio of volume to mole number is a constant.

BOYLE'S LAW: A statement, derived by English chemist Robert Boyle (1627-1691), which holds that for gases in isothermal conditions, an inverse relationship exists between the volume and pressure of a gas. This means that the greater the pressure, the less the volume and vice versa, and therefore the product of pressure multiplied by volume yields a constant figure.

CHARLES'S LAW: A statement, derived by French physicist and chemist J. A. C. Charles (1746-1823), which holds that for gases in isobarometric conditions, the ratio between the volume and temperature of a gas is constant. This means that the greater the temperature, the greater the volume and vice versa.

DALTON'S LAW OF PARTIAL PRESSURE: A statement, derived by the English chemist John Dalton (1766-1844), which holds that the total pressure of a gas is equal to the sum of its partial pres-

uations when the vehicle experiences extreme deceleration. When this occurs, there is a rapid transfer of kinetic energy to rest energy, as with the earlier illustration of a stone hitting concrete. And indeed, if you were to smash against a fully inflated airbag, it would feel like hitting concrete—with all the expected results.

The airbag's sensor contains a steel ball attached to a permanent magnet or a stiff spring. The spring holds it in place through minor mishaps in which an airbag would not be warranted—for instance, if a car were simply to be "tapped" by another in a parking lot. But in a case of sudden deceleration, the magnet or spring releases the ball, sending it down a smooth bore. It flips a switch, turning on an electrical circuit. This in turn ignites a pellet of sodium azide, which fills the bag with nitrogen gas.

The events described in the above illustration take place within 40 milliseconds—less time than it takes for your body to come flying forward; and then the airbag has to begin deflating before the body reaches it. At this point, the highly pressurized nitrogen gas molecules begin escaping through vents. Thus as your body hits the bag, the deflation of the latter is moving it in the same direction that your body is going—only much, much more slowly. Two seconds after impact, which is an eternity in terms of the processes involved, the pressure inside the bag has returned to 1 atm.

WHERE TO LEARN MORE

Beiser, Arthur. *Physics,* 5th ed. Reading, MA: Addison-Wesley, 1991.

"*Chemistry Units: Gas Laws.*" (Web site). <http://bio.bio.rpi.edu/MS99/ausemaW/chem/gases.hmtl> (February 21, 2001).

KEY TERMS CONTINUED

sures—that is, the pressure exerted by each component of the gas mixture.

GAY-LUSSAC'S LAW: A statement, derived by the French physicist and chemist Joseph Gay-Lussac (1778-1850), which holds that the pressure of a gas is directly related to its absolute temperature. Hence the ratio of pressure to absolute temperature is a constant.

HENRY'S LAW: A statement, derived by the English chemist William Henry (1774-836), which holds that the amount of gas dissolved in a liquid is directly proportional to the partial pressure of the gas above the solution. This holds true only for gases, such as hydrogen and oxygen, that are capable of dissolving in water without undergoing ionization.

IDEAL GAS LAW: A proposition, also known as the combined gas law, that draws on all the gas laws. The ideal gas law can be expressed as the formula $pV = nRT$, where p stands for pressure, V for volume, n for number of moles, and T for temperature. R is known as the universal gas constant, a figure equal to 0.0821 atm • liter/mole • K.

INVERSE RELATIONSHIP: A situation involving two variables, in which one of the two increases in direct proportion to the decrease in the other.

IONIZATION: A reaction in which an atom or group of atoms loses one or more electrons. The atoms are then converted to ions, which are either wholly positive or negative in charge.

ISOTHERMAL: Referring to a situation in which temperature is kept constant.

ISOBAROMETRIC: Referring to a situation in which pressure is kept constant.

MOLE: A unit equal to 6.022137×10^{23} molecules.

Laws of Gases. New York: Arno Press, 1981.

Macaulay, David. *The New Way Things Work.* Boston: Houghton Mifflin, 1998.

Mebane, Robert C. and Thomas R. Rybolt. *Air and Other Gases.* Illustrations by Anni Matsick. New York: Twenty-First Century Books, 1995.

"*Tutorials—6.*" <http://www.chemistrycoach.com/tutorials-6.html> (February 21, 2001).

MOLECULAR DYNAMICS

CONCEPT

Physicists study matter and motion, or matter in motion. These forms of matter may be large, or they may be far too small to be seen by the most high-powered microscopes available. Such is the realm of molecular dynamics, the study and simulation of molecular motion. As its name suggests, molecular dynamics brings in aspects of dynamics, the study of why objects move as they do, as well as thermodynamics, the study of the relationships between heat, work, and energy. Existing at the borders between physics and chemistry, molecular dynamics provides understanding regarding the properties of matter—including phenomena such as the liquefaction of gases, in which one phase of matter is transformed into another.

HOW IT WORKS

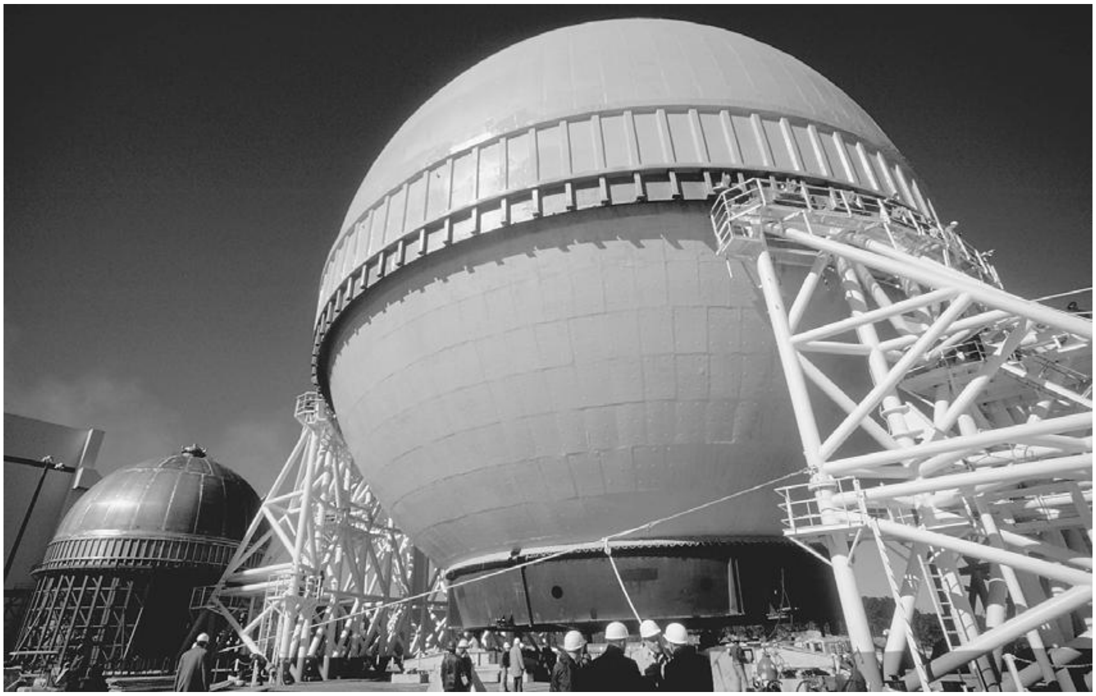
Molecules

The physical world is made up of matter, physical substance that has mass; occupies space; is composed of atoms; and is, ultimately, convertible to energy. On Earth, three principal phases of matter exist, namely solid, liquid, and gas. The differences between these three are, on the surface at least, easily perceivable. Clearly, water is a liquid, just as ice is a solid and steam a gas. Yet, the ways in which various substances convert between phases are often complex, as are the interrelations between these phases. Ultimately, understanding of the phases depends on an awareness of what takes place at the molecular level.

An atom is the smallest particle of a chemical element. It is not, however, the smallest thing in the universe; atoms are composed of subatomic particles, including protons, neutrons, and electrons. These subatomic particles are discussed in the context of the structure of matter elsewhere in this volume, where they are examined largely with regard to their electromagnetic properties. In the present context, the concern is primarily with the properties of atomic and molecular particles, in terms of mechanics, the study of bodies in motion, and thermodynamics.

An atom must, by definition, represent one and only one chemical element, of which 109 have been identified and named. It should be noted that the number of elements changes with continuing research, and that many of the elements, particularly those discovered relatively recently—as, for instance, meitnerium (No. 109), isolated in the 1990s—are hardly part of everyday experience. So, perhaps 100 would be a better approximation; in any case, consider the multitude of possible ways in which the elements can be combined.

Musicians have only seven tones at their disposal, and artists only seven colors—yet they manage to create a seemingly infinite variety of mutations in sound and sight, respectively. There are only 10 digits in the numerical system that has prevailed throughout the West since the late Middle Ages, yet it is possible to use that system to create such a range of numbers that all the books in all the libraries in the world could not contain them. This gives some idea of the range of combinations available using the hundred-odd chemical elements nature has provided—in other words, the number of possible molecular combinations that exist in the universe.



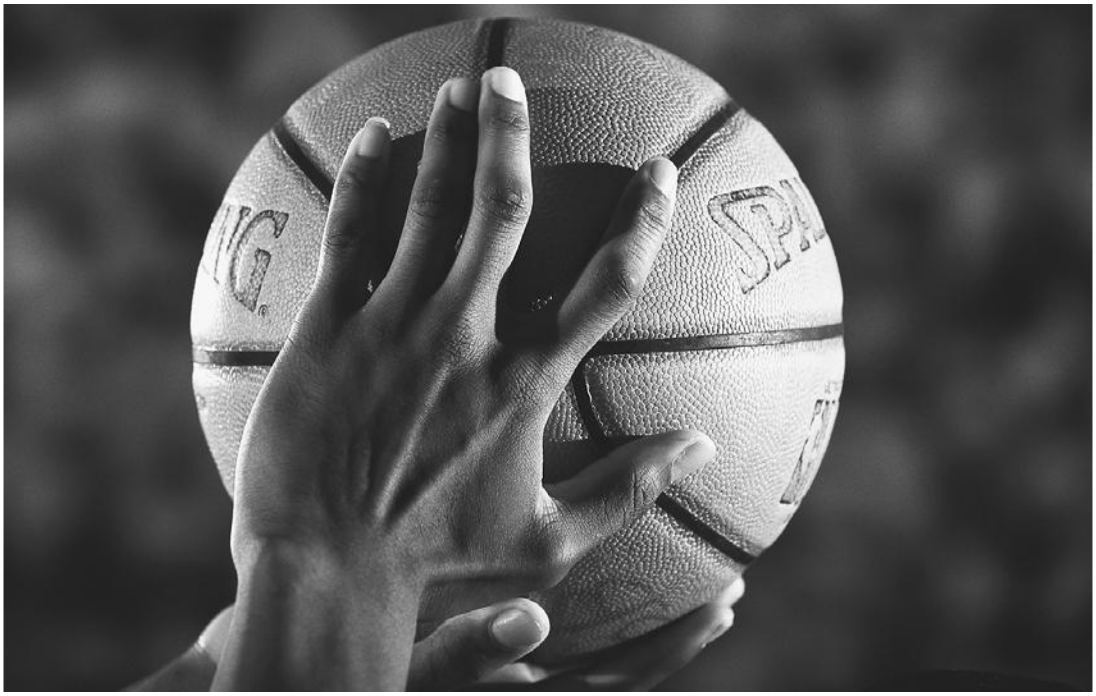
This huge liquefied natural gas container will be installed on a ship. The volume of the liquefied gas is far less than it would be if the gas were in a vaporized state, thus enabling ease and economy in transport. *(Photograph by James L. Amos/Corbis. Reproduced by permission.)*

THE STRUCTURE OF MOLECULES. A molecule is a group of atoms joined in a single structure. Often, these atoms come from different elements, in which case the molecule represents a particular chemical compound, such as water, carbon dioxide, sodium chloride (salt), and so on. On the other hand, a molecule may consist only of one type of atom: oxygen molecules, for instance, are formed by the joining of two oxygen atoms.

As much as scientists understand about molecules and their structure, there is much that they do not know. Molecules of water are fairly easy to understand, because they have a simple, regular structure that does not change. A water molecule is composed of one oxygen atom joined by two hydrogen atoms, and since the oxygen atom is much larger than the two hydrogens, its shape can be compared to a basketball with two softballs attached. The scale of the molecule, of course, is so small as to boggle the mind: to borrow an illustration from American physicist Richard Feynman (1918-1988), if a basketball were blown up to the size of Earth, the molecules inside of it would not even be as large as an ordinary-sized basketball.

As for the water molecule, scientists know a number of things about it: the distance between the two hydrogen atoms (measured in units called an angstrom), and even the angle at which they join the oxygen atom. In the case of salt, however, the molecular structure is not nearly as uniform as that of water: atoms join together, but not always in regular ways. And then there are compounds far more complex than water or salt, involving numerous elements that fit together in precise and complicated ways. But, once that discussion is opened, one has stepped from the realm of physics into that of chemistry, and that is not the intention here. Rather, the purpose of the foregoing and very cursory discussion of molecular structure is to point out that molecules are at the heart of all physical existence—and that the things we cannot see are every bit as complicated as those we can.

THE MOLE. Given the tiny—to use an understatement—size of molecules, how do scientists analyze their behavior? Today, physicists have at their disposal electron microscopes and other advanced forms of equipment that make it possible to observe activity at the atomic and molecular levels. The technology that makes this possible is beyond the scope of the present dis-



How small are molecules? If this basketball were blown up to the size of Earth, the molecules inside it would not be as big as a real basketball. *(Photograph by Dimitri Iundt/Corbis. Reproduced by permission.)*

cussion. On the other hand, consider a much simpler question: how do physicists weigh molecules?

Obviously "a bunch" of iron (an element known by the chemical symbol Fe) weighs more than "a bunch" of oxygen, but what exactly is "a bunch"? Italian physicist Amedeo Avogadro (1776-1856), the first scientist to clarify the distinction between atoms and molecules, created a unit that made it possible to compare the masses of various molecules. This is the mole, also known as "Avogadro's number," a unit equal to 6.022137×10^{23} (more than 600 billion trillion) molecules.

The term "mole" can be used in the same way that the word "dozen" is used. Just as "a dozen" can refer to twelve cakes or twelve chickens, so "mole" always describes the same number of molecules. A mole of any given substance has its own particular mass, expressed in grams. The mass of one mole of iron, for instance, will always be greater than that of one mole of oxygen. The ratio between them is exactly the same as the ratio of the mass of one iron atom to one oxygen atom. Thus, the mole makes it possible to compare the mass of one element or compound to that of another.

Molecular Attraction and Motion

Molecular dynamics can be understood primarily in terms of the principles of motion, identified by Sir Isaac Newton (1642-1727), principles that receive detailed discussion at several places in this volume. However, the attraction between particles at the atomic and molecular level cannot be explained by reference to gravitational force, also identified by Newton. For more than a century, gravity was the only type of force known to physicists, yet the pull of gravitation alone was too weak to account for the strong pull between atoms and molecules.

During the eighteenth century and early nineteenth centuries, however, physicists and other scientists became increasingly aware of another form of interaction at work in the world—one that could not be explained in gravitational terms. This was the force of electricity and magnetism, which Scottish physicist James Clerk Maxwell (1831-1879) suggested were different manifestations of a "new" kind of force, electromagnetism. All subatomic particles possess either a positive, negative, or neutral electrical charge. An atom usually has a neutral charge, meaning that it is composed of an equal number of protons (positive) and electrons (negative). In certain situations, however, it may lose one or more electrons and, thus, acquire a net charge, making it an ion.

Positive and negative charges attract one another, much as the north and south poles of two different magnets attract. (In fact, magnetism is simply an aspect of electromagnetic force.) Not only do the positive and negative elements of an atom attract one another, but positive elements in atoms attract negative elements in other atoms, and vice versa. These interactions are much more complex than the preceding discussion suggests, of course; the important point is that a force other than gravitation draws matter together at the atomic and molecular levels. On the other hand, the interactions that are critical to the study of molecular dynamics are primarily mechanical, comprehensible from the standpoint of Newtonian dynamics.

MOLECULAR BEHAVIOR AND PHASES OF MATTER. All molecules are in motion, and the rate of that motion is affected by the attraction between them. This attraction or repulsion can be though of like a spring connecting two molecules, an analogy that works best for solids, but in a limited way for liquids. Most molecular motion in liquids and gases is caused by collisions with other molecules; even in solids, momentum is transferred from one molecule to the next along the "springs," but ultimately the motion is caused by collisions. Hence molecular collisions provide the mechanism by which heat is transferred between two bodies in contact.

The rate at which molecules move in relation to one another determines phase of matter—that is, whether a particular item can be described as solid, liquid, or gas. The movement of molecules means that they possess kinetic energy, or the energy of movement, which is manifested as thermal energy and measured by temperature. Temperature is really nothing more than molecules in motion, relative to one another: the faster they move, the greater the kinetic energy, and the greater the temperature.

When the molecules in a material move slowly in relation to one another, they tend to be close in proximity, and hence the force of attraction between them is strong. Such a material is called a solid. In molecules of liquid, by contrast, the rate of relative motion is higher, so the molecules tend to be a little more spread out, and therefore the force between them is weaker. A material substance whose molecules move at high speeds, and therefore exert little attraction toward one another, is known as a gas. All forms of matter possess a certain (very large) amount of energy due to their mass; thermal energy, however, is—like phase of matter—a function of the attractions between particles. Hence, solids generally have less energy than liquids, and liquids less energy than gases.

REAL-LIFE APPLICATIONS

Kinetic Theories of Matter

English chemist John Dalton (1766-1844) was the first to recognize that nature is composed of tiny particles. In putting forward his idea, Dalton adopted a concept from the Greek philosopher Democritus (c. 470-380 B.C.), who proposed that matter is made up of tiny units he called atomos, or "indivisible."

Dalton recognized that the structure of atoms in a particular element or compound is uniform, and maintained that compounds are made up of compound atoms: in other words, that water, for instance, is composed of "water atoms." Soon after Dalton, however, Avogadro clarified the distinction between atoms and molecules. Neither Dalton nor Avogadro offered much in the way of a theory regarding atomic or molecular behavior; but another scientist had already introduced the idea that matter at the smallest levels is in a constant state of motion.

This was Daniel Bernoulli (1700-1782), a Swiss mathematician and physicist whose studies of fluids—a term which encompasses both gases and liquids—provided a foundation for the field of fluid mechanics. (Today, Bernoulli's principle, which relates the velocity and pressure of fluids, is applied in the field of aerodynamics, and explains what keeps an airplane aloft.) Bernoulli published his fluid mechanics studies in *Hydrodynamica* (1700-1782), a work in which he provided the basis for what came to be known as the kinetic theory of gases.

BROWNIAN MOTION. Because he came before Dalton and Avogadro, and, thus, did not have the benefit of their atomic and molecular theories, Bernoulli was not able to develop his kinetic theory beyond the seeds of an idea. The subsequent elaboration of kinetic theory, which is applied not only to gases but (with somewhat less effectiveness) to liquids and solids, in fact, resulted from an accidental discovery.

In 1827, Scottish botanist Robert Brown (1773-1858) was studying pollen grains under a microscope, when he noticed that the grains underwent a curious zigzagging motion in the water. The pollen assumed the shape of a colloid, a pattern that occurs when particles of one substance are dispersed—but not dissolved—in another substance. Another example of a colloidal pattern is a puff of smoke.

At first, Brown assumed that the motion had a biological explanation—that is, that it resulted from life processes within the pollen—but later, he discovered that even pollen from long-dead plants behaved in the same way. He never understood what he was witnessing. Nor did a number of other scientists, who began noticing other examples of what came to be known as Brownian motion: the constant but irregular zigzagging of colloidal particles, which can be seen clearly through a microscope.

MAXWELL, BOLTZMANN, AND THE MATURING OF KINETIC THEORY. A generation after Brown's time, kinetic theory came to maturity through the work of Maxwell and Austrian physicist Ludwig E. Boltzmann (1844-1906). Working independently, the two men developed a theory, later dubbed the Maxwell-Boltzmann theory of gases, which described the distribution of molecules in a gas. In 1859, Maxwell described the distribution of molecular velocities, work that became the foundation of statistical mechanics—the study of large systems—by examining the behavior of their smallest parts.

A year later, in 1860, Maxwell published a paper in which he presented the kinetic theory of gases: the idea that a gas consists of numerous molecules, relatively far apart in space, which interact by colliding. These collisions, he proposed, are responsible for the production of thermal energy, because when the velocity of the molecules increases—as it does after collision—the temperature increases as well. Eight years later, in 1868, Boltzmann independently applied statistics to the kinetic theory, explaining the behavior of gas molecules by means of what would come to be known as statistical mechanics.

Kinetic theory offered a convincing explanation of the processes involved in Brownian motion. According to the kinetic view, what Brown observed had nothing to do with the pollen particles; rather, the movement of those particles was simply the result of activity on the part of the water molecules. Pollen grains are many thousands of times larger than water molecules, but since there are so many molecules in even one drop of water, and their motion is so constant but apparently random, the water molecules are bound to move a pollen grain once every few thousand collisions.

In 1905, Albert Einstein (1879-1955) analyzed the behavior of particles subjected to Brownian motion. His work, and the confirmation of his results by French physicist Jean Baptiste Perrin (1870-1942), finally put an end to any remaining doubts concerning the molecular structure of matter. The kinetic explanation of molecular behavior, however, remains a theory.

KINETIC THEORY AND GASES

Maxwell's and Boltzmann's work helped explain characteristics of matter at the molecular level, but did so most successfully with regard to gases. Kinetic theory fits with a number of behaviors exhibited by gases: their tendency to fill any container by expanding to fit its interior, for instance, and their ability to be easily compressed.

This, in turn, concurs with the gas laws (discussed in a separate essay titled "Gas Laws")—for instance, Boyle's law, which maintains that pressure decreases as volume increases, and vice versa. Indeed, the ideal gas law, which shows an inverse relationship between pressure and volume, and a proportional relationship between temperature and the product of pressure and volume, is an expression of kinetic theory.

THE GAS LAWS ILLUSTRATED. The operations of the gas laws are easy to visualize by means of kinetic theory, which portrays gas molecules as though they were millions upon billions of tiny balls colliding at random. Inside a cube-shaped container of gas, molecules are colliding with every possible surface, but the net effect of these collisions is the same as though the molecules were divided into thirds, each third colliding with opposite walls inside the cube.

If the cube were doubled in size, the molecules bouncing back and forth between two sets of walls would have twice as far to travel between each collision. Their speed would not change, but the time between collisions would double, thus, cutting in half the amount of pressure they would exert on the walls. This is an illustration of Boyle's law: increasing the volume by a factor of two leads to a decrease in pressure to half of its original value.

On the other hand, if the size of the container were decreased, the molecules would have less distance to travel from collision to collision. This means they would be colliding with the walls more often, and, thus, would have a higher degree of energy—and, hence, a higher temperature. This illustrates another gas law, Charles's law, which relates volume to temperature: as one of the two increases or decreases, so does the other. Thus, it can be said, in light of kinetic theory, that the average kinetic energy produced by the motions of all the molecules in a gas is proportional to the absolute temperature of the gas.

GASES AND ABSOLUTE TEMPERATURE. The term "absolute temperature" refers to the Kelvin scale, established by William Thomson, Lord Kelvin (1824-1907). Drawing on Charles's discovery that gas at 0°C (32°F) regularly contracts by about 1/273 of its volume for every Celsius degree drop in temperature, Thomson derived the value of absolute zero (-273.15°C or -459.67°F). The Kelvin and Celsius scales are directly related; hence, Celsius temperatures can be converted to Kelvins by adding 273.15.

The Kelvin scale measures temperature in relation to absolute zero, or 0K. (Units in the Kelvin system, known as Kelvins, do not include the word or symbol for degree.) But what is absolute zero, other than a very cold temperature? Kinetic theory provides a useful definition: the temperature at which all molecular movement in a gas ceases. But this definition requires some qualification.

First of all, the laws of thermodynamics show the impossibility of actually reaching absolute zero. Second, the vibration of atoms never completely ceases: rather, the vibration of the average atom is zero. Finally, one element—helium—does not freeze, even at temperatures near absolute zero. Only the application of pressure will push helium past the freezing point.

Changes of Phase

Kinetic theory is more successful when applied to gases than to liquids and solids, because liquid and solid molecules do not interact nearly as frequently as gas particles do. Nonetheless, the proposition that the internal energy of any substance—gas, liquid, or solid—is at least partly related to the kinetic energies of its molecules helps explain much about the behavior of matter.

The thermal expansion of a solid, for instance, can be clearly explained in terms of kinetic theory. As discussed in the essay on elasticity, many solids are composed of crystals, regular shapes composed of molecules joined to one another, as though on springs. A spring that is pulled back, just before it is released, is an example of potential energy: the energy that an object possesses by virtue of its position. For a crystalline solid at room temperature, potential energy and spacing between molecules are relatively low. But as temperature increases and the solid expands, the space between molecules increases—as does the potential energy in the solid.

An example of a liquid displaying kinetic behavior is water in the process of vaporization. The vaporization of water, of course, occurs in boiling, but water need not be anywhere near the boiling point to evaporate. In either case, the process is the same. Speeds of molecules in any substance are distributed along a curve, meaning that a certain number of molecules have speeds well below, or well above, the average. Those whose speeds are well above the average have enough energy to escape the surface, and once they depart, the average energy of the remaining liquid is less than before. As a result, evaporation leads to cooling. (In boiling, of course, the continued application of thermal energy to the entire water sample will cause more molecules to achieve greater energy, even as highly energized molecules leave the surface of the boiling water as steam.)

The Phase Diagram

The vaporization of water is an example of a change of phase—the transition from one phase of matter to another. The properties of any substance, and the points at which it changes phase, are plotted on what is known as a phase diagram. The latter typically shows temperature along the x-axis, and pressure along the y-axis. It is also possible to construct a phase diagram that plots volume against temperature, or volume against pressure, and there are even three-dimensional phase diagrams that measure the relationship between all three—volume, pressure, and temperature. Here we will consider the simpler two-dimensional diagram we have described.

For simple substances such as water and carbon dioxide, the solid form of the substance appears at a relatively low temperature, and at pressures anywhere from zero upward. The line between solids and liquids, indicating the temperature at which a solid becomes a liquid at any pressure above a certain level, is called the fusion curve. Though it appears to be a line, it is indeed curved, reflecting the fact that at high pressures, a solid well below the normal freezing point for that substance may be melted to create a liquid.

Liquids occupy the area of the phase diagram corresponding to relatively high temperatures and high pressures. Gases or vapors, on the other hand, can exist at very low temperatures, but only if the pressure is also low. Above the melting point for the substance, gases exist at higher pressures and higher temperatures. Thus, the line between liquids and gases often looks almost like a 45° angle. But it is not a straight line, as its name, the vaporization curve, indicates. The curve of vaporization reflects the fact that at relatively high temperatures and high pressures, a substance is more likely to be a gas than a liquid.

CRITICAL POINT AND SUBLIMATION. There are several other interesting phenomena mapped on a phase diagram. One is the critical point, which can be found at a place of very high temperature and pressure along the vaporization curve. At the critical point, high temperatures prevent a liquid from remaining a liquid, no matter how high the pressure. At the same time, the pressure causes gas beyond that point to become more and more dense, but due to the high temperatures, it does not condense into a liquid. Beyond the critical point, the substance cannot exist in anything other than the gaseous state. The temperature component of the critical point for water is 705.2°F (374°C)—at 218 atm, or 218 times ordinary atmospheric pressure. For helium, however, critical temperature is just a few degrees above absolute zero. This is why helium is rarely seen in forms other than a gas.

There is also a certain temperature and pressure, called the triple point, at which some substances—water and carbon dioxide are examples—will be a liquid, solid, and gas all at once. Another interesting phenomenon is the sublimation curve, or the line between solid and gas. At certain very low temperatures and pressures, a substance may experience sublimation, meaning that a gas turns into a solid, or a solid into a gas, without passing through a liquid stage. A well-known example of sublimation occurs when "dry ice," which is made of carbon dioxide, vaporizes at temperatures above (-109.3°F [-78.5°C]). Carbon dioxide is exceptional, however, in that it experiences sublimation at relatively high pressures, such as those experienced in everyday life: for most substances, the sublimation point occurs at such a low pressure point that it is seldom witnessed outside of a laboratory.

Liquefaction of Gases

One interesting and useful application of phase change is the liquefaction of gases, or the change of gas into liquid by the reduction in its molecular energy levels. There are two important properties at work in liquefaction: critical temperature and critical pressure. Critical temperature is that temperature above which no amount of pressure will cause a gas to liquefy. Critical pressure is the amount of pressure required to liquefy the gas at critical temperature.

Gases are liquefied by one of three methods: (1) application of pressure at temperatures below critical; (2) causing the gas to do work against external force, thus, removing its energy and changing it to the liquid state; or (3) causing the gas to do work against some internal force. The second option can be explained in terms of the operation of a heat engine, as explored in the Thermodynamics essay.

In a steam engine, an example of a heat engine, water is boiled, producing energy in the form of steam. The steam is introduced to a cylinder, in which it pushes on a piston to drive some type of machinery. In pushing against the piston, the steam loses energy, and as a result, changes from a gas back to a liquid.

As for the use of internal forces to cool a gas, this can be done by forcing the vapor through a small nozzle or porous plug. Depending on the temperature and properties of the gas, such an

KEY TERMS

ABSOLUTE ZERO: The temperature, defined as 0K on the Kelvin scale, at which the motion of molecules in a solid virtually ceases. Absolute zero is equal to -459.67°F (-273.15°C).

ATOM: The smallest particle of a chemical element. An atom can exist either alone or in combination with other atoms in a molecule.

BROWNIAN MOTION: The constant but irregular zigzagging of colloidal particles, which can be seen clearly through a microscope. The phenomenon is named after Scottish botanist Robert Brown (1773-1858), who first witnessed it but was not able to explain it. The behavior exhibited in Brownian motion provides evidence for the kinetic theory of matter.

CHANGE OF PHASE: The transition from one phase of matter to another.

CHEMICAL COMPOUND: A substance made up of atoms of more than one chemical element. These atoms are usually joined in molecules.

CHEMICAL ELEMENT: A substance made up of only one kind of atom.

COLLOID: A pattern that occurs when particles of one substance are dispersed—but not dissolved—in another substance. A puff of smoke in the air is an example of a colloid, whose behavior is typically characterized by Brownian motion.

CRITICAL POINT: A coordinate, plotted on a phase diagram, above which a substance cannot exist in anything other than the gaseous state. Located at a position of very high temperature and pressure, the critical point marks the termination of the vaporization curve.

DYNAMICS: The study of why objects move as they do. Dynamics is an element of mechanics.

FLUID: Any substance, whether gas or liquid, which tends to flow, and which conforms to the shape of its container. Unlike solids, fluids are typically uniform in molecular structure: for instance, one molecule of water is the same as another water molecule.

FUSION CURVE: The boundary between solid and liquid for any given substance, as plotted on a phase diagram.

GAS: A phase of matter in which molecules exert little or no attraction toward one another, and, therefore, move at high speeds.

HEAT: Internal thermal energy that flows from one body of matter to another.

KELVIN SCALE: Established by William Thomson, Lord Kelvin (1824-1907), the Kelvin scale measures temperature in relation to absolute zero, or 0K. (Units in the Kelvin system, known as Kelvins, do not include the word or symbol for degree.) The Kelvin and Celsius scales are directly related; hence, Celsius temperatures can be converted to Kelvins by adding 273.15.

KINETIC ENERGY: The energy that an object possesses by virtue of its motion.

KINETIC THEORY OF GASES: The idea that a gas consists of numerous molecules, relatively far apart in space, which interact by colliding. These collisions are responsible for the production of thermal energy, because when the velocity of the molecules increases—as it does after collision—the temperature increases as well.

KEY TERMS CONTINUED

KINETIC THEORY OF MATTER: The application of the kinetic theory of gases to all forms of matter. Since particles of liquids and solids move much more slowly than do gas particles, kinetic theory is not as successful in this regard; however, the proposition that the internal energy of any substance is at least partly related to the kinetic energies of its molecules helps explain much about the behavior of matter.

LIQUID: A phase of matter in which molecules exert moderate attractions toward one another, and, therefore, move at moderate speeds.

MATTER: Physical substance that has mass; occupies space; is composed of atoms; and is ultimately convertible to energy. There are several phases of matter, including solids, liquids, and gases.

MECHANICS: The study of bodies in motion.

MOLE: A unit equal to 6.022137×10^{23} (more than 600 billion trillion) molecules. Since their size makes it impossible to weigh molecules in relatively small quantities; hence, the mole, devised by Italian physicist Amedeo Avogadro (1776-1856), facilitates comparisons of mass between substances.

MOLECULAR DYNAMICS: The study and simulation of molecular motion.

MOLECULE: A group of atoms, usually of more than one chemical element, joined in a structure.

PHASE DIAGRAM: A chart, plotted for any particular substance, identifying the particular phase of matter for that substance at a given temperature and pressure level. A phase diagram usually shows temperature along the x-axis, and pressure along the y-axis.

PHASES OF MATTER: The various forms of material substance (matter),

operation may be enough to remove energy sufficient for liquefaction to take place. Sometimes, the process must be repeated before the gas fully condenses into a liquid.

HISTORICAL BACKGROUND. Like the steam engine itself, the idea of gas liquefaction is a product of the early Industrial Age. One of the pioneering figures in the field was the brilliant English physicist Michael Faraday (1791-1867), who liquefied a number of high-critical temperature gases, such as carbon dioxide.

Half a century after Faraday, French physicist Louis Paul Cailletet (1832-1913) and Swiss chemist Raoul Pierre Pictet (1846-1929) developed the nozzle and porous-plug methods of liquefaction. This, in turn, made it possible to liquefy gases with much lower critical temperatures, among them oxygen, nitrogen, and carbon monoxide.

By the end of the nineteenth century, physicists were able to liquefy the gases with the lowest critical temperatures. James Dewar of Scotland (1842-1923) liquefied hydrogen, whose critical temperature is -399.5°F (-239.7°C). Some time later, Dutch physicist Heike Kamerlingh Onnes (1853-1926) successfully liquefied the gas with the lowest critical temperature of them all: helium, which, as mentioned earlier, becomes a gas at almost unbelievably low temperatures. Its critical temperature is -449.9°F (-267.7°C), or just 5.3K.

APPLICATIONS OF GAS LIQUEFACTION. Liquefied natural gas (LNG)

KEY TERMS CONTINUED

which are defined primarily in terms of the behavior exhibited by their atomic or molecular structures. On Earth, three principal phases of matter exist, namely solid, liquid, and gas.

POTENTIAL ENERGY: The energy an object possesses by virtue of its position.

SOLID: A phase of matter in which molecules exert strong attractions toward one another, and, therefore, move slowly.

STATISTICAL MECHANICS: A realm of the physical sciences devoted to the study of large systems by examining the behavior of their smallest parts.

SUBLIMATION CURVE: The boundary between solid and gas for any given substance, as plotted on a phase diagram.

SYSTEM: In physics, the term "system" usually refers to any set of physical interactions isolated from the rest of the universe. Anything outside of the system, including all factors and forces irrelevant to a discussion of that system, is known as the environment.

TEMPERATURE: A measure of the average kinetic energy—or molecular translational energy in a system. Differences in temperature determine the direction of internal energy flow between two systems when heat is being transferred.

THERMAL ENERGY: Heat energy, a form of kinetic energy produced by the movement of atomic or molecular particles. The greater the movement of these particles, the greater the thermal energy.

THERMODYNAMICS: The study of the relationships between heat, work, and energy.

VAPORIZATION CURVE: The boundary between liquid and gas for any given substance as plotted on a phase diagram.

and liquefied petroleum gas (LPG), the latter a mixture of by-products obtained from petroleum and natural gas, are among the examples of liquefied gas in daily use. In both cases, the volume of the liquefied gas is far less than it would be if the gas were in a vaporized state, thus enabling ease and economy in transport.

Liquefied gases are used as heating fuel for motor homes, boats, and homes or cabins in remote areas. Other applications of liquefied gases include liquefied oxygen and hydrogen in rocket engines, and liquefied oxygen and petroleum used in welding. The properties of liquefied gases also figure heavily in the science of producing and studying low-temperature environments. In addition, liquefied helium is used in studying the behavior of matter at temperatures close to absolute zero.

A "New" Form of Matter?

Physicists at a Colorado laboratory in 1995 revealed a highly interesting aspect of atomic motion at temperatures approaching absolute zero. Some 70 years before, Einstein had predicted that, at extremely low temperatures, atoms would fuse to form one large "superatom." This hypothesized structure was dubbed the Bose-Einstein Condensate after Einstein and Satyendranath Bose (1894-1974), an Indian physicist whose statistical methods contributed to the development of quantum theory.

Because of its unique atomic structure, the Bose-Einstein Condensate has been dubbed a "new" form of matter. It represents a quantum mechanical effect, relating to a cutting-edge area of physics devoted to studying the properties of subatomic particles and the interaction of matter with radiation. Thus it is not directly related to molecular dynamics; nonetheless, the Bose-Einstein Condensate is mentioned here as an example of the exciting work being performed at a level beyond that addressed by molecular dynamics. Its existence may lead to a greater understanding of quantum mechanics, and on an everyday level, the "superatom" may aid in the design of smaller, more powerful computer chips.

WHERE TO LEARN MORE

Cooper, Christopher. *Matter.* New York: DK Publishing, 1999.

"Kinetic Theory of Gases: A Brief Review" University of Virginia Department of Physics (Web site). <http://www.phys.virginia.edu/classes/252/kinetic_theory.html> (April 15, 2001).

"*The Kinetic Theory Page*" (Web site). <http://comp.uark.edu/~jgeabana/mol_dyn/> (April 15, 2001).

Medoff, Sol and John Powers. *The Student Chemist Explores Atoms and Molecules.* Illustrated by Nancy Lou Gahan. New York: R. Rosen Press, 1977.

"*Molecular Dynamics*" (Web site). <http://www.biochem.vt.edu/courses/modeling/molecular_dynamics.html> (April 15, 2001).

"*Molecular Simulation Molecular Dynamics Page*" (Web site). <http://www.phy.bris.ac.uk/research/theory/simulation/md.html> (April 15, 2001).

Santrey, Laurence. *Heat.* Illustrated by Lloyd Birmingham. Mahwah, NJ: Troll Associates, 1985.

Strasser, Ben. *Molecules in Motion.* Illustrated by Vern Jorgenson. Pasadena, CA: Franklin Publications, 1967.

Van, Jon. "*U.S. Scientists Create a 'Superatom.'*" *Chicago Tribune,* July 14, 1995, p. 3.

STRUCTURE OF MATTER

CONCEPT

The physical realm is made up of matter. On Earth, matter appears in three clearly defined forms—solid, liquid, and gas—whose visible and perceptible structure is a function of behavior that takes place at the molecular level. Though these are often referred to as "states" of matter, it is also useful to think of them as phases of matter. This terminology serves as a reminder that any one substance can exist in any of the three phases. Water, for instance, can be ice, liquid, or steam; given the proper temperature and pressure, it may be solid, liquid, and gas all at once! But the three definite earthbound states of matter are not the sum total of the material world: in outer space a fourth phase, plasma, exists—and there may be still other varieties in the physical universe.

HOW IT WORKS

Matter and Energy

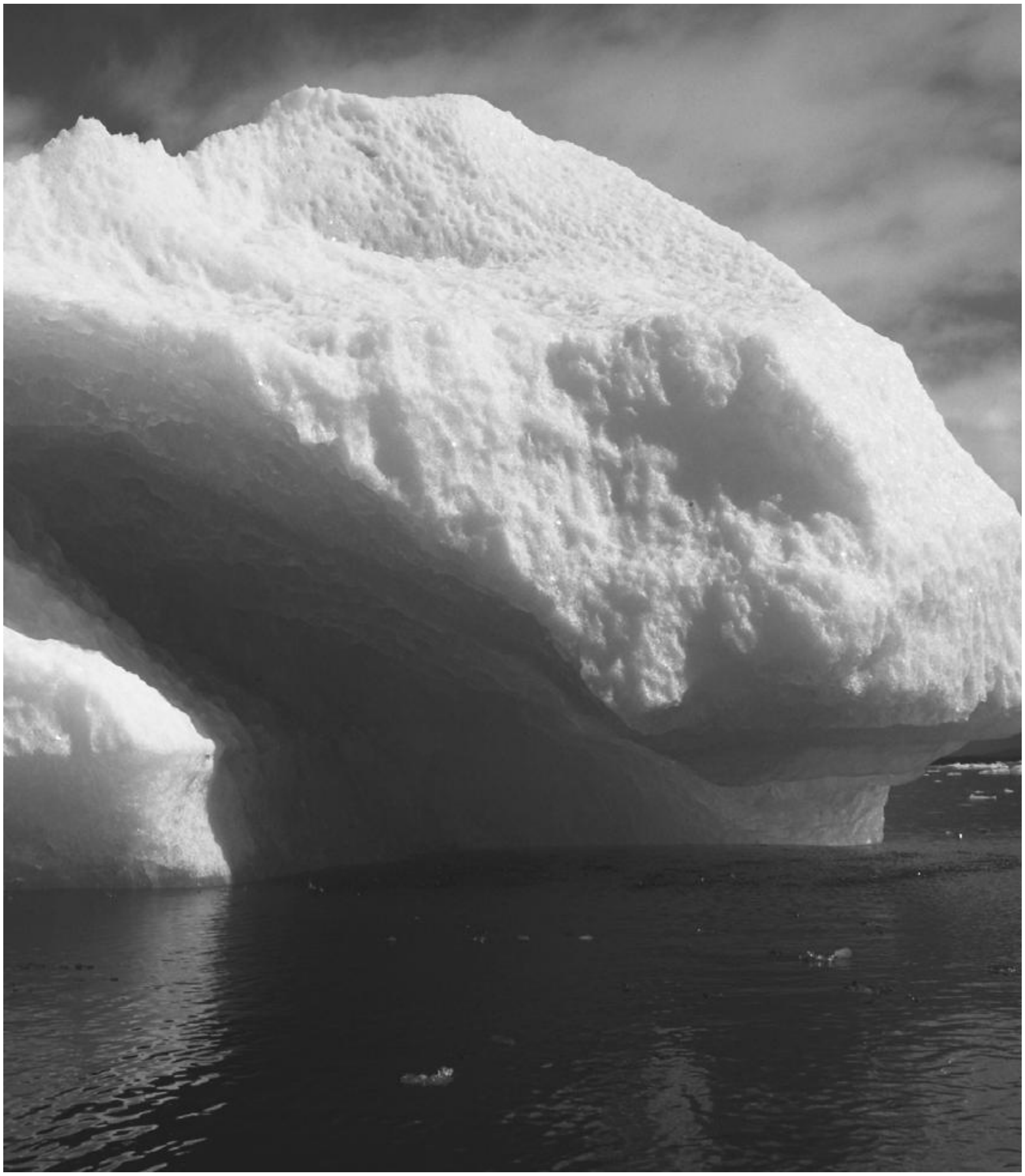
Matter can be defined as physical substance that has mass; occupies space; is composed of atoms; and is ultimately convertible to energy. A significant conversion of matter to energy, however, occurs only at speeds approaching that of the speed of light, a fact encompassed in the famous statement formulated by Albert Einstein (1879-1955), $E = mc^2$.

Einstein's formula means that every item possesses a quantity of energy equal to its mass multiplied by the squared speed of light. Given the fact that light travels at 186,000 mi (297,600 km) per second, the quantities of energy available from even a tiny object traveling at that speed are massive indeed. This is the basis for both nuclear power and nuclear weaponry, each of which uses some of the smallest particles in the known universe to produce results that are both amazing and terrifying.

The forms of matter that most people experience in their everyday lives, of course, are traveling at speeds well below that of the speed of light. Even so, transfers between matter and energy take place, though on a much, much smaller scale. For instance, when a fire burns, only a tiny fraction of its mass is converted to energy. The rest is converted into forms of mass different from that of the wood used to make the fire. Much of it remains in place as ash, of course, but an enormous volume is released into the atmosphere as a gas so filled with energy that it generates not only heat but light. The actual mass converted into energy, however, is infinitesimal.

CONSERVATION AND CONVERSION. The property of energy is, at all times and at all places in the physical universe, conserved. In physics, "to conserve" something means "to result in no net loss of" that particular component—in this case, energy. Energy is never destroyed: it simply changes form. Hence, the conservation of energy, a law of physics stating that within a system isolated from all other outside factors, the total amount of energy remains the same, though transformations of energy from one form to another take place.

Whereas energy is perfectly conserved, matter is only approximately conserved, as shown with the example of the fire. Most of the matter from the wood did indeed turn into more mat-



AN ICEBERG FLOATS BECAUSE THE DENSITY OF ICE IS LOWER THAN WATER, WHILE ITS VOLUME IS GREATER, MAKING THE ICEBERG BUOYANT. *(Photograph by Ric Engenbright/Corbis. Reproduced by permission.)*

ter—that is, vapor and ash. Yet, as also noted, a tiny quantity of matter—too small to be perceived by the senses—turned into energy.

The conservation of mass holds that total mass is constant, and is unaffected by factors such as position, velocity, or temperature, in any system that does not exchange any matter with its environment. This, however, is a qualified statement: at speeds well below c (the speed of light), it is essentially true, but for matter approaching c and thus, turning into energy, it is not.

Consider an item of matter moving at the speed of 100 mi (160 km)/sec. This is equal to 360,000 MPH (576,000 km/h) and in terms of the speeds to which humans are accustomed, it seems incredibly fast. After all, the fastest any human beings have ever traveled was about 25,000 MPH (40,000 km/h), in the case of the astronauts aboard *Apollo 11* in May 1969, and the speed under discussion is more than 14 times greater. Yet 100 mi/sec is a snail's pace compared to c: in fact, the proportional difference between an actual snail's pace and the speed of a human walking is not as great. Yet even at this leisurely gait, equal to 0.00054*c,* a portion of mass equal to 0.0001% (one-millionth of the total mass) converts to energy.

Matter at the Atomic Level

In his brilliant work *Six Easy Pieces,* American physicist Richard Feynman (1918-1988) asked his readers, "If, in some cataclysm, all of scientific knowledge were to be destroyed, and only one sentence passed on to the next generations of creatures, what statement would contain the most information in the fewest words? I believe it is the atomic hypothesis (or the atomic fact, or whatever you wish to call it) that all things are made of atoms—little articles that move around in perpetual motion, attracting each other when they are a little distance apart, but repelling upon being squeezed into one another. In that sentence, you will see, there is an enormous amount of information about the world, if just a little imagination and thinking are applied."

Feynman went on to offer a powerful series of illustrations concerning the size of atoms relative to more familiar objects: if an apple were magnified to the size of Earth, for instance, the atoms in it would each be about the size of a regular apple. Clearly atoms and other atomic particles are far too small to be glimpsed by even the most highly powered optical microscope. Yet, it is the behavior of particles at the atomic level that defines the shape of the entire physical world. Viewed from this perspective, it becomes easy to understand how and why matter is convertible to energy. Likewise, the interaction between atoms and other particles explains why some types of matter are solid, others liquid, and still others, gas.

ATOMS AND MOLECULES. An atom is the smallest particle of a chemical element. It is not, however, the smallest particle in the universe; atoms are composed of subatomic particles, including protons, neutrons, and electrons. But at the subatomic level, it is meaningless to refer to, for instance, "an oxygen electron": electrons are just electrons. An atom, then, is the fundamental unit of matter. Most of the substances people encounter in the world, however, are not pure elements, such as oxygen or iron; they are chemical compounds, in which atoms of more than one element join together to form molecules.

One of the most well-known molecular forms in the world is water, or H_2O, composed of two hydrogen atoms and one oxygen atom. The arrangement is extremely precise and never varies: scientists know, for instance, that the two hydrogen atoms join the oxygen atom (which is much larger than the hydrogen atoms) at an angle of 105° 3'. Other molecules are much more complex than those of water—some of them much, much more complex, which is reflected in the sometimes unwieldy names required to identify their chemical components.

ATOMIC AND MOLECULAR THEORY. The idea of atoms is not new. More than 24 centuries ago, the Greek philosopher Democritus (c. 470-380 B.C.) proposed that matter is composed of tiny particles he called *atomos,* or "indivisible." Democritus was not, however, describing matter in a concrete, scientific way: his "atoms" were idealized, philosophical constructs, not purely physical units.

Yet, he came amazingly close to identifying the fundamental structure of physical reality—much closer than any number of erroneous theories (such as the "four elements" of earth, air, fire, and water) that prevailed until modern times. English chemist John Dalton (1766-1844) was the first to identify what Feynman later called the "atomic hypothesis": that nature is composed of tiny particles. In putting forward his idea, Dalton adopted Democritus's word "atom" to describe these basic units.

Dalton recognized that the structure of atoms in a particular element or compound is uniform. He maintained that compounds are made up of compound atoms: in other words, that water, for instance, is a compound of "water atoms." Water, however, is not an element, and thus, it was necessary to think of its atomic composition in a different way—in terms of molecules rather than atoms. Dalton's contemporary Amedeo Avogadro (1776-1856), an Italian physicist, was the first scientist to clarify the distinction between atoms and molecules.

THE MOLE. Obviously, it is impractical to weigh a single molecule, or even several thousand; what was needed, then, was a number large enough to make possible practical comparisons of mass. Hence, the mole, a quantity equal to "Avogadro's number." The latter, named after Avogadro though not derived by him, is equal to 6.022137×10^{23} (more than 600 billion trillion) molecules.

The term "mole" can be used in the same way that the word "dozen" is used. Just as "a dozen" can refer to twelve cakes or twelve chickens, so "mole" always describes the same number of molecules. A mole of any given substance has its own particular mass, expressed in grams. The mass of one mole of iron, for instance, will always be greater than that of one mole of oxygen. The ratio between them is exactly the same as the ratio of the mass of one iron atom to one oxygen atom. Thus, the mole makes it possible to compare the mass of one element or compound to that of another.

BROWNIAN MOTION AND KINETIC THEORY. Contemporary to both Dalton and Avogadro was Scottish naturalist Robert Brown (1773-1858), who in 1827 stumbled upon a curious phenomenon. While studying pollen grains under a microscope, Brown noticed that the grains underwent a curious zigzagging motion in the water. At first, he assumed that the motion had a biological explanation—that is, it resulted from life processes within the pollen—but later he discovered that even pollen from long-dead plants behaved in the same way.

Brown never understood what he was witnessing. Nor did a number of other scientists, who began noticing other examples of what came to be known as Brownian motion: the constant but irregular zigzagging of particles in a puff of smoke, for instance. Later, however, Scottish physicist James Clerk Maxwell (1831-1879) and others were able to explain it by what came to be known as the kinetic theory of matter.

The kinetic theory, which is discussed in depth elsewhere in this book, is based on the idea that molecules are constantly in motion: hence, the water molecules were moving the pollen grains Brown observed. Pollen grains are many thousands of times as large as water molecules, but there are so many molecules in just one drop of water, and their motion is so constant but apparently random, that they are bound to move a pollen grain once every few thousand collisions.

GROWTH IN UNDERSTANDING THE ATOM. Einstein, who was born the year Maxwell died, published a series of papers in which he analyzed the behavior of particles subjected to Brownian motion. His work, and the confirmation of his results by French physicist Jean Baptiste Perrin (1870-1942), finally put an end to any remaining doubts concerning the molecular structure of matter.

It may seem amazing that the molecular and atomic ideas were still open to question in the early twentieth century; however, the vast majority of what is known today concerning the atom emerged after World War I. At the end of the nineteenth century, scientists believed the atom to be indivisible, but growing evidence concerning electrical charges in atoms brought with it the awareness that there must be something smaller creating those charges.

Eventually, physicists identified protons and electrons, but the neutron, with no electrical charge, was harder to discover: it was not identified until 1932. After that point, scientists were convinced that just three types of subatomic particles existed. However, subsequent activity among physicists—particularly those in the field of quantum mechanics—led to the discovery of other elementary particles, such as the photon. However, in this discussion, the only subatomic particles whose behavior is reviewed are the proton, electron, and neutron.

Motion and Attraction in Atoms and Molecules

At the molecular level, every item of matter in the world is in motion. This may be easy enough to imagine with regard to air or water, since both tend to flow. But what about a piece of paper, or a glass, or a rock? In fact, all molecules are in constant motion, and depending on the particular phase of matter, this motion may vary from a mere vibration to a high rate of speed.

Molecular motion generates kinetic energy, or the energy of movement, which is manifested as heat or thermal energy. Indeed, heat is really nothing more than molecules in motion relative to one another: the faster they move, the greater the kinetic energy, and the greater the heat.

The movement of atoms and molecules is always in a straight line and at a constant velocity, unless acted upon by some outside force. In fact, the motion of atoms and molecules is constantly being interfered with by outside forces, because they are perpetually striking one another. These collisions cause changes in direction, and may lead to transfers of energy from one particle to another.

ELECTROMAGNETIC FORCE IN ATOMS. The behavior of molecules cannot be explained in terms of gravitational force. This force, and the motions associated with it, were identified by Sir Isaac Newton (1642-1727), and Newton's model of the universe seemed to answer most physical questions. Then in the late nineteenth century, Maxwell discovered a second kind of force, electromagnetism. (There are two other known varieties of force, strong and weak nuclear, which are exhibited at the subatomic level.) Electromagnetic force, rather than gravitation, explains the attraction between atoms.

Several times up to this point, the subatomic particles have been mentioned but not explained in terms of their electrical charge, which is principal among their defining characteristics. Protons have a positive electrical charge, while neutrons exert no charge. These two types of particles, which make up the vast majority of the atom's mass, are clustered at the center, or nucleus. Orbiting around this nucleus are electrons, much smaller particles which exert a negative charge.

Chemical elements are identified by the number of protons they possess. Hydrogen, first element listed on the periodic table of elements, has one proton and is thus identified as 1; carbon, or element 6, has six protons, and so on.

An atom usually has a neutral charge, meaning that it is composed of an equal number of protons or electrons. In certain situations, however, it may lose one or more electrons and thus acquire a net charge. Such an atom is called an ion. But electrical charge, like energy, is conserved, and the electrons are not "lost" when an atom becomes an ion: they simply go elsewhere.

MOLECULAR BEHAVIOR AND STATES OF MATTER. Positive and negative charges interact at the molecular level in a way that can be compared to the behavior of poles in a pair of magnets. Just as two north poles or two south poles repel one another, so like charges—two positives, or two negatives—repel. Conversely, positive and negative charges exert an attractive force on one another similar to that of a north pole and south pole in contact.

In discussing phases of matter, the attraction between molecules provides a key to distinguishing between states of matter. This is not to say one particular phase of matter is a particularly good conductor of electrical current, however. For instance, certain solids—particularly metals such as copper—are extremely good conductors. But wood is a solid, too, and conducts electrical current poorly.

The properties of various forms of matter, viewed from the larger electromagnetic picture, are a subject far beyond the scope of this essay. In any case, the electromagnetic properties of concern in the present instance are not the ones demonstrated at a macroscopic level—that is, in view of "the big picture." Rather, the subject of the attractive force operating at the atomic or molecular levels has been introduced to show that certain types of material have a greater intermolecular attraction.

As previously stated, all matter is in motion. The relative speed of that motion, however, is a function of the attraction between molecules, which in turn defines a material according to one of the phases of matter. When the molecules in a material exert a strong attraction toward one another, they move slowly, and the material is called a solid. Molecules of liquid, by contrast, exert a moderate attraction and move at moderate speeds. A material substance whose molecules exert little or no attraction, and therefore, move at high speeds, is known as a gas.

These comparisons of molecular speed and attraction, obviously, are relative. Certainly, it is easy enough in most cases to distinguish between one phase of matter and another, but there are some instances in which they overlap. Examples of these will follow, but first it is necessary to discuss the phases of matter in the context of their behavior in everyday situations.

REAL-LIFE APPLICATIONS

From Solid to Liquid

The attractions between particles have a number of consequences in defining the phases of matter. The strong attractive forces in solids cause its particles to be positioned close together. This means that particles of solids resist attempts to compress them, or push them together. Because of their close proximity, solid particles are fixed in an orderly and definite pattern. As a result, a solid usually has a definite volume and shape.

A crystal is a type of solid in which the constituent parts are arranged in a simple, definite geometric arrangement that is repeated in all directions. Metals, for instance, are crystalline solids. Other solids are said to be amorphous, meaning that they possess no definite shape. Amorphous solids—clay, for example—either possess very tiny crystals, or consist of several varieties of crystal mixed randomly. Still other solids, among them glass, do not contain crystals.

VIBRATIONS AND FREEZING. Because of their strong attractions to one another, solid particles move slowly, but like all particles of matter, they do move. Whereas the particles in a liquid or gas move fast enough to be in relative motion with regard to one another, however, solid particles merely vibrate from a fixed position.

This can be shown by the example of a singer hitting a certain note and shattering a glass. Contrary to popular belief, the note does not have to be particularly high: rather, the note should be on the same wavelength as the vibration of the glass. When this occurs, sound energy is transferred directly to the glass, which shatters because of the sudden net intake of energy.

As noted earlier, the attraction and motion of particles in matter has a direct effect on heat and temperature. The cooler the solid, the slower and weaker the vibrations, and the closer the particles are to one another. Thus, most types of matter contract when freezing, and their density increases. Absolute zero, or 0K on the Kelvin scale of temperature—equal to -459.67°F (-273°C)—is the point at which vibration virtually stops. Note that the vibration virtually stops, but does not stop entirely. In any event, the lowest temperature actually achieved, at a Finnish nuclear laboratory in 1993, is $2.8 \cdot 10^{-10}$K, or 0.00000000028K—still above absolute zero.

UNUSUAL CHARACTERISTICS OF ICE. The behavior of water at the freezing/melting point is interesting and exceptional. Above 39.2°F (4°C) water, like most substances, expands when heated. But between 32°F (0°C) and that temperature, however, it actually contracts. And whereas most substances become much denser with lowered temperatures, the density of water reaches its maximum at 39.2°F. Below that point, it starts to decrease again.

Not only does the density of ice begin decreasing just before freezing, but its volume increases. This is the reason ice floats: its weight is less than that of the water it has displaced, and therefore, it is buoyant. Additionally, the buoyant qualities of ice atop very cold water explain why the top of a lake may freeze, but lakes rarely freeze solid—even in the coldest of inhabited regions.

Instead of freezing from the bottom up, as it would if ice were less buoyant than the water, the lake freezes from the top down. Furthermore, ice is a poorer conductor of heat than water, and, thus, little of the heat from the water below escapes. Therefore, the lake does not freeze completely—only a layer at the top—and this helps preserve animal and plant life in the body of water. On the other hand, the increased volume of frozen water is not always good for humans: when water in pipes freezes, it may increase in volume to the point where the pipe bursts.

MELTING. When heated, particles begin to vibrate more and more, and, therefore, move further apart. If a solid is heated enough, it loses its rigid structure and becomes a liquid. The temperature at which a solid turns into a liquid is called the melting point, and melting points are different for different substances. For the most part, however, solids composed of heavier particles require more energy—and, hence, higher temperatures—to induce the vibrations necessary for freezing. Nitrogen melts at -346°F (-210°C), ice at 32°F (0°C), and copper at 1,985°F (1,085°C). The melting point of a substance, incidentally, is the same as its freezing point: the difference is a matter of orientation—that is, whether the process is one of a solid melting to become a liquid, or of a liquid freezing to become a solid.

The energy required to change a solid to a liquid is called the heat of fusion. In melting, all the heat energy in a solid (energy that exists due to the motion of its particles) is used in breaking up the arrangement of crystals, called a lattice. This is why the water resulting from melted ice does not feel any warmer than when it was frozen: the thermal energy has been expended, with none left over for heating the water. Once all the ice is melted, however, the absorbed energy from the particles—now moving at much greater speeds than when the ice was in a solid state—causes the temperature to rise.

From Liquid to Gas

The particles of a liquid, as compared to those of a solid, have more energy, more motion, and less attraction to one another. The attraction, however, is still fairly strong: thus, liquid particles are in close enough proximity that the liquid resists compression.

On the other hand, their arrangement is loose enough that the particles tend to move around one another rather than merely vibrating in place, as solid particles do. A liquid is therefore not definite in shape. Both liquids and gases tend to flow, and to conform to the shape of their container; for this reason, they are together classified as fluids.

Owing to the fact that the particles in a liquid are not as close in proximity as those of a solid, liquids tend to be less dense than solids. The liquid phase of substance is thus inclined to be larger in volume than its equivalent in solid form. Again, however, water is exceptional in this regard: liquid water actually takes up less space than an equal mass of frozen water.

BOILING. When a liquid experiences an increase in temperature, its particles take on energy and begin to move faster and faster. They collide with one another, and at some point the particles nearest the surface of the liquid acquire enough energy to break away from their neighbors. It is at this point that the liquid becomes a gas or vapor.

As heating continues, particles throughout the liquid begin to gain energy and move faster, but they do not immediately transform into gas. The reason is that the pressure of the liquid, combined with the pressure of the atmosphere above the liquid, tends to keep particles in place. Those particles below the surface, therefore, remain where they are until they acquire enough energy to rise to the surface.

The heated particle moves upward, leaving behind it a hollow space—a bubble. A bubble is not an empty space: it contains smaller trapped particles, but its small weight relative to that of the liquid it disperses makes it buoyant. Therefore, a bubble floats to the top, releasing its trapped particles as gas or vapor. At that point, the liquid is said to be boiling.

THE EFFECT OF ATMOSPHERIC PRESSURE. As they rise, the particles thus have to overcome atmospheric pressure, and this means that the boiling point for any liquid depends in part on the pressure of the surrounding air. This is why cooking instructions often vary with altitude: the greater the distance from sea level, the less the air pressure, and the shorter the required cooking time.

Atop Mt. Everest, Earth's highest peak at about 29,000 ft (8,839 m) above sea level, the pressure is approximately one-third normal atmospheric pressure. This means the air is one-third as dense as it is as sea level, which explains why mountain-climbers on Everest and other tall peaks must wear oxygen masks to stay alive. It also means that water boils at a much lower temperature on Everest than it does elsewhere. At sea level, the boiling point of water is 212°F (100°C), but at 29,000 ft it is reduced by one-quarter, to 158°F (70°C).

Of course, no one lives on the top of Mt. Everest—but people do live in Denver, Colorado, where the altitude is 5,577 ft (1,700 m) and the boiling point of water is 203°F (95°C). Given the lower boiling point, one might assume that food would cook faster in Denver than in New York, Los Angeles, or some other city close to sea level. In fact, the opposite is true: because heated particles escape the water so much faster at high altitudes, they do not have time to acquire the energy needed to raise the temperature of the water. It is for this reason that a recipe may contain a statement such as "at altitudes above XX feet, add XX minutes to cooking time."

If lowered atmospheric pressure means a lowered boiling point, what happens in outer space, where there is no atmospheric pressure? Liquids boil at very, very low temperatures. This is one of the reasons why astronauts have to wear pressurized suits: if they did not, their blood would boil—even though space itself is incredibly cold.

LIQUID TO GAS AND BACK AGAIN. Note that the process of a liquid changing to a gas is similar to what occurs when a solid changes to a liquid: particles gain heat and therefore energy, begin to move faster, break free from one another, and pass a certain threshold into a new phase of matter. And just as the freezing and melting point for a given substance are the same temperature, the only difference being one of orientation, the boiling point of a liquid transforming into a gas is the same as the condensation point for a gas turning into a liquid.

The behavior of water in boiling and condensation makes possible distillation, one of the principal methods for purifying seawater in various parts of the world. First, the water is boiled, then, it is allowed to cool and condense, thus forming water again. In the process, the water separates from the salt, leaving it behind in the form of brine. A similar separation takes place when salt water freezes: because salt, like most solids, has a much lower freezing point than water, very little of it remains joined to the water in ice. Instead, the salt takes the form of a briny slush.

GAS AND ITS LAWS. Having reached the gaseous state, a substance takes on characteristics quite different from those of a solid, and somewhat different from those of a liquid. Whereas liquid particles exert a moderate attraction to one another, particles in a gas exert little to no attraction. They are thus free to move, and to move quickly. The shape and arrangement of gas is therefore random and indefinite—and, more importantly, the motion of gas particles give it much greater kinetic energy than the other forms of matter found on Earth.

The constant, fast, and random motion of gas particles means that they are always colliding and thereby transferring kinetic energy back and forth without any net loss. These collisions also have the overall effect of producing uniform pressure in a gas. At the same time, the characteristics and behavior of gas particles indicate that they will tend not to remain in an open container. Therefore, in order to maintain any pressure on a gas—other than the normal atmospheric pressure exerted on the surface of the gas by the atmosphere (which, of course, is also a gas)—it is necessary to keep it in a closed container.

There are a number of gas laws (examined in another essay in this book) describing the response of gases to changes in pressure, temperature, and volume. Among these is Boyle's law, which holds that when the temperature of a gas is constant, there is an inverse relationship between volume and pressure: in other words, the greater the pressure, the less the volume, and vice versa. According to a second gas law, Charles's law, for gases in conditions of constant pressure, the ratio between volume and temperature is constant—that is, the greater the temperature, the greater the volume, and vice versa.

In addition, Gay-Lussac's law shows that the pressure of a gas is directly related to its absolute temperature on the Kelvin scale: the higher the temperature, the higher the pressure, and vice versa. Gay-Lussac's law is combined, along with Boyle's and Charles's and other gas laws, in the ideal gas law, which makes it possible to find the value of any one variable—pressure, volume, number of moles, or temperature—for a gas, as long as one knows the value of the other three.

Other States of Matter

PLASMA. Principal among states of matter other than solid, liquid, and gas is plasma, which is similar to gas. (The term "plasma," when referring to the state of matter, has nothing to do with the word as it is often used, in reference to blood plasma.) As with gas, plasma particles collide at high speeds—but in plasma, the speeds are even greater, and the kinetic energy levels even higher.

The speed and energy of these collisions is directly related to the underlying property that distinguishes plasma from gas. So violent are the collisions between plasma particles that electrons are knocked away from their atoms. As a result, plasma does not have the atomic structure typical of a gas; rather, it is composed of positive ions and electrons. Plasma particles are thus electrically charged, and, therefore, greatly influenced by electrical and magnetic fields.

Formed at very high temperatures, plasma is found in stars and comets' tails; furthermore, the reaction between plasma and atomic particles in the upper atmosphere is responsible for the aurora borealis or "northern lights." Though not found on Earth, plasma—ubiquitous in other parts of the universe—may be the most plentiful among the four principal states of matter.

QUASI-STATES. Among the quasi-states of matter discussed by physicists are several terms that describe the structure in which particles are joined, rather than the attraction and relative movement of those particles. "Crystalline," "amorphous," and "glassy" are all terms to describe what may be individual states of matter; so too is "colloidal."

A colloid is a structure intermediate in size between a molecule and a visible particle, and it has a tendency to be dispersed in another medium—as smoke, for instance, is dispersed in air. Brownian motion describes the behavior of most colloidal particles. When one sees dust floating in a ray of sunshine through a window, the light reflects off colloids in the dust, which are driven back and forth by motion in the air otherwise imperceptible to the human senses.

DARK MATTER. The number of states or phases of matter is clearly not fixed, and it is quite possible that more will be discovered in outer space, if not on Earth. One intriguing candidate is called dark matter, so described because it neither reflects nor emits light, and is therefore invisible. In fact, luminous or visible matter may very well make up only a small fraction of the mass in the universe, with the rest being taken up by dark matter.

If dark matter is invisible, how do astronomers and physicists know it exists? By analyzing the gravitational force exerted on visible objects when there seems to be no visible object to account for that force. An example is the center of our galaxy, the Milky Way, which appears to be nothing more than a dark "halo." In order to cause the entire galaxy to revolve around it in the same way that planets revolve around the Sun, the Milky Way must contain a staggering quantity of invisible mass.

Dark matter may be the substance at the heart of a black hole, a collapsed star whose mass is so great that its gravitational field prevents light from escaping. It is possible, also, that dark matter is made up of neutrinos, subatomic particles thought to be massless. Perhaps, the theory goes, neutrinos actually possess tiny quantities of mass, and therefore in huge groups—a mole times a mole times a mole—they might possess appreciable mass.

In addition, dark matter may be the deciding factor as to whether the universe is infinite. The more mass the universe possesses, the greater its overall gravity, and if the mass of the universe is above a certain point, it will eventually begin to contract. This, of course, would mean that it is finite; on the other hand, if the mass is below this threshold, it will continue to expand indefinitely. The known mass of the universe is nowhere near that threshold—but, because the nature of dark matter is still largely unknown, it is not possible yet to say what effect its mass may have on the total equation.

A "NEW" FORM OF MATTER? Physicists at the Joint Institute of Laboratory Astrophysics in Boulder, Colorado, in 1995 revealed a highly interesting aspect of atomic behavior at temperatures approaching absolute zero. Some 70 years before, Einstein had predicted that, at extremely low temperatures, atoms would fuse to form one large "superatom." This hypothesized structure was dubbed the Bose-Einstein Condensate (BEC) after Einstein and Satyendranath Bose (1894-1974), an Indian physicist whose statistical methods contributed to the development of quantum theory.

Cooling about 2,000 atoms of the element rubidium to a temperature just 170 billionths of a degree Celsius above absolute zero, the physicists succeeded in creating an atom 100 micrometers across—still incredibly small, but vast in comparison to an ordinary atom. The superatom, which lasted for about 15 seconds, cooled down all the way to just 20 billionths of a degree above absolute zero. The Colorado physicists won the Nobel Prize in physics in 1997 for their work.

In 1999, researchers in a lab at Harvard University also created a superatom of BEC, and used it to slow light to just 38 MPH (60.8 km/h)—about 0.02% of its ordinary speed. Dubbed a "new" form of matter, the BEC may lead to a greater understanding of quantum mechanics, and may aid in the design of smaller, more powerful computer chips.

States and Phases and In Between

At places throughout this essay, references have been made variously to "phases" and "states" of matter. This is not intended to confuse, but rather to emphasize a particular point. Solids, liquids, and gases are referred to as "phases," because substances on Earth—water, for instance—regularly move from one phase to another. This change, a function of temperature, is called (aptly enough) "change of phase."

There is absolutely nothing incorrect in referring to "states of matter." But "phases of matter" is used in the present context as a means of emphasizing the fact that most substances, at the appropriate temperature and pressure, can be solid, liquid, or gas. In fact, a substance may even be solid, liquid, and gas.

An Analogy to Human Life

The phases of matter can be likened to the phases of a person's life: infancy, babyhood, childhood, adolescence, adulthood, old age. The transition between these stages is indefinite, yet it is

The response of liquid crystals to light makes them useful in the displays used on laptop computers. *(AFP/Corbis. Reproduced by permission.)*

easy enough to say when a person is at a certain stage.

At the transition point between adolescence and adulthood—say, at seventeen years old—a young person may say that she is an adult, but her parents may insist that she is still an adolescent or a child. And indeed, she might qualify as either. On the other hand, when she is thirty, it would be ridiculous to assert that she is anything other than an adult.

At the same time, a person at a certain age may exhibit behaviors typically associated with another age. A child, for instance, may behave like an adult, or an adult like a baby. One interesting example of this is the relationship between age two and late adolescence. In both cases, the person is in the process of individualizing, developing an identity separate from that of his or her parents—yet clearly, there are also plenty of differences between a two-year-old and a seventeen-year-old.

As with the transitional phases in human life, in the borderline pressure levels and temperatures for phases of matter it is sometimes difficult to say, for instance, if a substance is fully a liquid or fully a gas. On the other hand, at a certain temperature and pressure level, a substance clearly is what it is: water at very low temperature and pressure, for instance, is indisputably ice—just as an average thirty-year-old is obviously an adult. As for the second observation, that a person at one stage in life may reflect characteristics of another stage, this too is reflected in the behavior of matter.

LIQUID CRYSTALS. A liquid crystal is a substance that, over a specific range of temperature, displays properties both of a liquid and a solid. Below this temperature range, it is unquestionably a solid, and above this range it is just as obviously a liquid. In between, however, liquid crystals exhibit a strange solid-liquid behavior: like a liquid, their particles flow, but like a solid, their molecules maintain specific crystalline arrangements.

Long, wide, and placed alongside one another, liquid crystal molecules exhibit interesting properties in response to light waves. The speed of light through a liquid crystal actually varies, depending on whether the light is traveling along the short or long sides of the molecules. These differences in light speed may lead to a change in the direction of polarization, or the vibration of light waves.

KEY TERMS

ATOM: The smallest particle of a chemical element. An atom can exist either alone or in combination with other atoms in a molecule. Atoms are made up of protons, neutrons, and electrons. In most cases, the electrical charges in atoms cancel out one another; but when an atom loses one or more electrons, and thus has a net charge, it becomes an ion.

CHEMICAL COMPOUND: A substance made up of atoms of more than one chemical element. These atoms are usually joined in molecules.

CHEMICAL ELEMENT: A substance made up of only one kind of atom.

CONSERVATION OF ENERGY: A law of physics which holds that within a system isolated from all other outside factors, the total amount of energy remains the same, though transformations of energy from one form to another take place.

CONSERVATION OF MASS: A physical principle which states that total mass is constant, and is unaffected by factors such as position, velocity, or temperature, in any system that does not exchange any matter with its environment. Unlike the other conservation laws, however, conservation of mass is not universally applicable, but applies only at speeds significant lower than that of light—186,000 mi (297,600 km) per second. Close to the speed of light, mass begins converting to energy.

CONSERVE: In physics, "to conserve" something means "to result in no net loss of" that particular component. It is possible that within a given system, the component may change form or position, but as long as the net value of the component remains the same, it has been conserved.

ELECTRON: Negatively charged particles in an atom. Electrons, which spin around the nucleus of protons and neutrons, constitute a very small portion of the atom's mass. In most atoms, the number of electrons and protons is the same, thus canceling out one another. When an atom loses one or more electrons, however—thus becoming an ion—it acquires a net electrical charge.

FRICTION: The force that resists motion when the surface of one object comes into contact with the surface of another.

FLUID: Any substance, whether gas or liquid, that tends to flow, and that conforms to the shape of its container. Unlike solids, fluids are typically uniform in molecular structure for instance, one molecule of water is the same as another water molecule.

GAS: A phase of matter in which molecules exert little or no attraction toward one another, and therefore move at high speeds.

ION: An atom that has lost or gained one or more electrons, and thus has a net electrical charge.

LIQUID: A phase of matter in which molecules exert moderate attractions toward one another, and therefore move at moderate speeds.

MATTER: Physical substance that has mass; occupies space; is composed of atoms; and is ultimately (at speeds approaching that of light) convertible to energy. There are several phases of matter, including solids, liquids, and gases.

KEY TERMS CONTINUED

MOLE: A unit equal to 6.022137×10^{23} (more than 600 billion trillion) molecules. Their size makes it impossible to weigh molecules in relatively small quantities; hence the mole facilitates comparisons of mass between substances.

MOLECULE: A group of atoms, usually of more than one chemical element, joined in a structure.

NEUTRON: A subatomic particle that has no electrical charge. Neutrons are found at the nucleus of an atom, alongside protons.

PHASES OF MATTER: The various forms of material substance (matter), which are defined primarily in terms of the behavior exhibited by their atomic or molecular structures. On Earth, three principal phases of matter exist, namely solid, liquid, and gas. Other forms of matter include plasma.

PLASMA: One of the phases of matter, closely related to gas. Plasma apparently does not exist on Earth, but is found, for instance, in stars and comets' tails. Containing neither atoms nor molecules, plasma is made up of electrons and positive ions.

PROTON: A positively charged particle in an atom. Protons and neutrons, which together form the nucleus around which electrons orbit, have approximately the same mass—a mass that is many times greater than that of an electron.

SOLID: A phase of matter in which molecules exert strong attractions toward one another, and therefore move slowly.

SYSTEM: In physics, the term "system" usually refers to any set of physical interactions isolated from the rest of the universe. Anything outside of the system, including all factors and forces irrelevant to a discussion of that system, is known as the environment.

The cholesteric class of liquid crystals is so named because the spiral patterns of light through the crystal are similar to those which appear in cholesterols. Depending on the physical properties of a cholesteric liquid crystal, only certain colors may be reflected. The response of liquid crystals to light makes them useful in liquid crystal displays (LCDs) found on laptop computer screens, camcorder views, and in other applications.

In some cholesteric liquid crystals, high temperatures lead to a reflection of shorter visible light waves, and lower temperatures to a display of longer visible waves. Liquid crystal thermometers thus show red when cool, and blue as they are warmed. This may seem a bit unusual to someone who does not understand why the thermometer displays those colors, since people typically associate red with heat and blue with cold.

THE TRIPLE POINT. A liquid crystal exhibits aspects of both liquid and solid, and thus, at certain temperatures may be classified within the crystalline quasi-state of matter. On the other hand, the phenomenon known as the triple point shows how an ordinary substance, such as water or carbon dioxide, can actually be a liquid, solid, and vapor—all at once.

Again, water—the basis of all life on Earth—is an unusual substance in many regards. For instance, most people associate water as a gas or vapor (that is, steam) with very high temperatures. Yet, at a level far below normal atmospheric pressure, water can be a vapor at temperatures as low as -4°F (-20 °C). (All of the pressure values in the discussion of water at or near the triple point are far below atmospheric norms: the pressure at which water would turn into a vapor at -4°F, for instance, is about 1/1000 normal atmospheric pressure.)

As everyone knows, at relatively low temperatures, water is a solid—ice. But if the pressure of ice falls below a very low threshold, it will turn straight into a gas (a process known as sublimation) without passing through the liquid stage. On the other hand, by applying enough pressure, it is possible to melt ice, and thereby transform it from a solid to a liquid, at temperatures below its normal freezing point.

The phases and changes of phase for a given substance at specific temperatures and pressure levels can be plotted on a graph called a phase diagram, which typically shows temperature on the x-axis and pressure on the y-axis. The phase diagram of water shows a line between the solid and liquid states that is almost, but not quite, exactly perpendicular to the x-axis: it slopes slightly upward to the left, reflecting the fact that solid ice turns into water with an increase of pressure.

Whereas the line between solid and liquid water is more or less straight, the division between these two states and water vapor is curved. And where the solid-liquid line intersects the vaporization curve, there is a place called the triple point. Just below freezing, in conditions equivalent to about 0.7% of normal atmospheric pressure, water is a solid, liquid, and vapor all at once.

WHERE TO LEARN MORE

Biel, Timothy L. *Atom: Building Blocks of Matter.* San Diego, CA: Lucent Books, 1990.

Feynman, Richard. *Six Easy Pieces: Essentials of Physics Explained by Its Most Brilliant Teacher.* New introduction by Paul Davies. Cambridge, MA: Perseus Books, 1995.

Hewitt, Sally. *Solid, Liquid, or Gas?* New York: Children's Press, 1998.

"High School Chemistry Table of Contents—Solids and Liquids" Homeworkhelp.com (Web site). <http://www.homeworkhelp.com/homeworkhelp/freemember/text/chem/hig h/topic09.html> (April 10, 2001).

"Matter: Solids, Liquids, Gases." Studyweb (Web site). <http://www.studyweb.com/links/4880.html> (April 10, 2001).

"The Molecular Circus" (Web site). <http://www.cpo.com/Weblabs/circus.html> (April 10, 2001).

Paul, Richard. *A Handbook to the Universe: Explorations of Matter, Energy, Space, and Time for Beginning Scientific Thinkers.* Chicago: Chicago Review Press, 1993.

"Phases of Matter" (Web site). <http://pc65.frontier.osrhe.edu/hs/science/pphase.html> (April 10, 2001).

Royston, Angela. *Solids, Liquids, and Gasses.* Chicago: Heinemann Library, 2001.

Wheeler, Jill C. *The Stuff Life's Made Of: A Book About Matter.* Minneapolis, MN: Abdo & Daughters Publishing, 1996.

THERMODYNAMICS

CONCEPT

Thermodynamics is the study of the relationships between heat, work, and energy. Though rooted in physics, it has a clear application to chemistry, biology, and other sciences: in a sense, physical life itself can be described as a continual thermodynamic cycle of transformations between heat and energy. But these transformations are never perfectly efficient, as the second law of thermodynamics shows. Nor is it possible to get "something for nothing," as the first law of thermodynamics demonstrates: the work output of a system can never be greater than the net energy input. These laws disappointed hopeful industrialists of the early nineteenth century, many of whom believed it might be possible to create a perpetual motion machine. Yet the laws of thermodynamics did make possible such highly useful creations as the internal combustion engine and the refrigerator.

HOW IT WORKS

Historical Context

Machines were, by definition, the focal point of the Industrial Revolution, which began in England during the late eighteenth and early nineteenth centuries. One of the central preoccupations of both scientists and industrialists thus became the efficiency of those machines: the ratio of output to input. The more output that could be produced with a given input, the greater the production, and the greater the economic advantage to the industrialists and (presumably) society as a whole.

At that time, scientists and captains of industry still believed in the possibility of a perpetual motion machine: a device that, upon receiving an initial input of energy, would continue to operate indefinitely without further input. As it emerged that work could be converted into heat, a form of energy, it began to seem possible that heat could be converted directly back into work, thus making possible the operation of a perfectly reversible perpetual motion machine. Unfortunately, the laws of thermodynamics dashed all those dreams.

SNOW'S EXPLANATION. Some texts identify two laws of thermodynamics, while others add a third. For these laws, which will be discussed in detail below, British writer and scientist C. P. Snow (1905-1980) offered a witty, nontechnical explanation. In a 1959 lecture published as *The Two Cultures and the Scientific Revolution,* Snow compared the effort to transform heat into energy, and energy back into heat again, as a sort of game.

The first law of thermodynamics, in Snow's version, teaches that the game is impossible to win. Because energy is conserved, and thus, its quantities throughout the universe are always the same, one cannot get "something for nothing" by extracting more energy than one put into a machine.

The second law, as Snow explained it, offers an even more gloomy prognosis: not only is it impossible to win in the game of energy-work exchanges, one cannot so much as break even. Though energy is conserved, that does not mean the energy is conserved within the machine where it is used: mechanical systems tend toward increasing disorder, and therefore, it is impossi-

A WOMAN WITH A SUNBURNED NOSE. SUNBURNS ARE CAUSED BY THE SUN'S ULTRAVIOLET RAYS. *(Photograph by Lester V. Bergman/Corbis. Reproduced by permission.)*

ble for the machine even to return to the original level of energy.

The third law, discovered in 1905, seems to offer a possibility of escape from the conditions imposed in the second law: at the temperature of absolute zero, this tendency toward breakdown drops to a level of zero as well. But the third law only proves that absolute zero cannot be attained: hence, Snow's third observation, that it is impossible to step outside the boundaries of this unwinnable heat-energy transformation game.

WORK AND ENERGY

Work and energy, discussed at length elsewhere in this volume, are closely related. Work is the exertion of force over a given distance to displace or move an object. It is thus the product of force and distance exerted in the same direction. Energy is the ability to accomplish work.

There are many manifestations of energy, including one of principal concern in the present context: thermal or heat energy. Other manifestations include electromagnetic (sometimes divided into electrical and magnetic), sound, chemical, and nuclear energy. All these, however, can be described in terms of mechanical energy, which is the sum of potential energy—the energy that an object has due to its position—and kinetic energy, or the energy an object possesses by virtue of its motion.

MECHANICAL ENERGY. Kinetic energy relates to heat more clearly than does potential energy, discussed below; however, it is hard to discuss the one without the other. To use a simple example—one involving mechanical energy in a gravitational field—when a stone is held over the edge of a cliff, it has potential energy. Its potential energy is equal to its weight (mass times the acceleration due to gravity) multiplied by its height above the bottom of the canyon below. Once it is dropped, it acquires kinetic energy, which is the same as one-half its mass multiplied by the square of its velocity.

Just before it hits bottom, the stone's kinetic energy will be at a maximum, and its potential energy will be at a minimum. At no point can the value of its kinetic energy exceed the value of the potential energy it possessed before it fell: the mechanical energy, or the sum of kinetic and potential energy, will always be the same, though the relative values of kinetic and potential energy may change.

CONSERVATION OF ENERGY. What mechanical energy does the stone possess after it comes to rest at the bottom of the canyon? In terms of the system of the stone dropping from the cliffside to the bottom, none. Or, to put it another way, the stone has just as much mechanical energy as it did at the very beginning. Before it was picked up and held over the side of the cliff, thus giving it potential energy, it was presumably sitting on the ground away from the edge of the cliff. Therefore, it lacked potential energy, inasmuch as it could not be "dropped" from the ground.

If the stone's mechanical energy—at least in relation to the system of height between the cliff and the bottom—has dropped to zero, where did it go? A number of places. When it hit, the stone transferred energy to the ground, manifested as heat. It also made a sound when it landed, and this also used up some of its energy. The stone itself lost energy, but the total energy in the universe was unaffected: the energy simply left the stone and went to other places. This is an example of the conservation of energy, which is closely tied to the first law of thermodynamics.

But does the stone possess any energy at the bottom of the canyon? Absolutely. For one thing, its mass gives it an energy, known as mass or rest energy, that dwarfs the mechanical energy in the system of the stone dropping off the cliff. (Mass energy is the other major form of energy, aside from kinetic and potential, but at speeds well below that of light, it is released in quantities that are virtually negligible.) The stone may have electromagnetic potential energy as well; and of course, if someone picks it up again, it will have gravitational potential energy. Most important to the present discussion, however, is its internal kinetic energy, the result of vibration among the molecules inside the stone.

Heat and Temperature

Thermal energy, or the energy of heat, is really a form of kinetic energy between particles at the atomic or molecular level: the greater the movement of these particles, the greater the thermal energy. Heat itself is internal thermal energy that flows from one body of matter to another. It is not the same as the energy contained in a system—that is, the internal thermal energy of the system. Rather than being "energy-in-residence," heat is "energy-in-transit."

This may be a little hard to comprehend, but it can be explained in terms of the stone-and-cliff kinetic energy illustration used above. Just as a system can have no kinetic energy unless something is moving within it, heat exists only when energy is being transferred. In the above illustration of mechanical energy, when the stone was sitting on the ground at the top of the cliff, it was analogous to a particle of internal energy in body *A*. When, at the end, it was again on the ground—only this time at the bottom of the canyon—it was the same as a particle of internal energy that has transferred to body *B*. In between, however, as it was falling from one to the other, it was equivalent to a unit of heat.

TEMPERATURE. In everyday life, people think they know what temperature is: a measure of heat and cold. This is wrong for two reasons: first, as discussed below, there is no such thing as "cold"—only an absence of heat. So, then, is temperature a measure of heat? Wrong again.

Imagine two objects, one of mass *M* and the other with a mass twice as great, or 2*M*. Both have a certain temperature, and the question is, how much heat will be required to raise their temperature by equal amounts? The answer is that the object of mass 2*M* requires twice as much heat to raise its temperature the same amount. Therefore, temperature cannot possibly be a measure of heat.

What temperature does indicate is the direction of internal energy flow between bodies, and the average molecular kinetic energy in transit between those bodies. More simply, though a bit less precisely, it can be defined as a measure of heat differences. (As for the means by which a thermometer indicates temperature, that is beyond the parameters of the subject at hand; it is discussed elsewhere in this volume, in the context of thermal expansion.)

MEASURING TEMPERATURE AND HEAT. Temperature, of course, can be measured either by the Fahrenheit or Centigrade scales familiar in everyday life. Another temperature scale of relevance to the present discussion is the Kelvin scale, established by William Thomson, Lord Kelvin (1824-1907).

Drawing on the discovery made by French physicist and chemist J. A. C. Charles (1746-

1823), that gas at 0°C (32°F) regularly contracts by about 1/273 of its volume for every Celsius degree drop in temperature, Thomson derived the value of absolute zero (discussed below) as -273.15°C (-459.67°F). The Kelvin and Celsius scales are thus directly related: Celsius temperatures can be converted to Kelvins (for which neither the word nor the symbol for "degree" are used) by adding 273.15.

MEASURING HEAT AND HEAT CAPACITY. Heat, on the other hand, is measured not by degrees (discussed along with the thermometer in the context of thermal expansion), but by the same units as work. Since energy is the ability to perform work, heat or work units are also units of energy. The principal unit of energy in the SI or metric system is the joule (J), equal to 1 newton-meter (N • m), and the primary unit in the British or English system is the foot-pound (ft • lb). One foot-pound is equal to 1.356 J, and 1 joule is equal to 0.7376 ft • lb.

Two other units are frequently used for heat as well. In the British system, there is the Btu, or British thermal unit, equal to 778 ft • lb. or 1,054 J. Btus are often used in reference, for instance, to the capacity of an air conditioner. An SI unit that is also used in the United States—where British measures typically still prevail—is the kilocalorie. This is equal to the heat that must be added to or removed from 1 kilogram of water to change its temperature by 1°C. As its name suggests, a kilocalorie is 1,000 calories. A calorie is the heat required to change the temperature in 1 gram of water by 1°C—but the dietary Calorie (capital C), with which most people are familiar is the same as the kilocalorie.

A kilocalorie is identical to the heat capacity for one kilogram of water. Heat capacity (sometimes called specific heat capacity or specific heat) is the amount of heat that must be added to, or removed from, a unit of mass for a given substance to change its temperature by 1°C. this is measured in units of J/kg • °C (joules per kilogram-degree Centigrade), though for the sake of convenience it is typically rendered in terms of kilojoules (1,000 joules): kJ/kg • °c. Expressed thus, the specific heat of water 4.185—which is fitting, since a kilocalorie is equal to 4.185 kJ. Water is unique in many aspects, with regard to specific heat, in that it requires far more heat to raise the temperature of water than that of mercury or iron.

REAL-LIFE APPLICATIONS

HOT AND "COLD"

Earlier, it was stated that there is no such thing as "cold"—a statement hard to believe for someone who happens to be in Buffalo, New York, or International Falls, Minnesota, during a February blizzard. Certainly, cold is real as a sensory experience, but in physical terms, cold is not a "thing"—it is simply the absence of heat.

People will say, for instance, that they put an ice cube in a cup of coffee to cool it, but in terms of physics, this description is backward: what actually happens is that heat flows from the coffee to the ice, thus raising its temperature. The resulting temperature is somewhere between that of the ice cube and the coffee, but one cannot obtain the value simply by averaging the two temperatures at the beginning of the transfer.

For one thing, the volume of the water in the ice cube is presumably less than that of the water in the coffee, not to mention the fact that their differing chemical properties may have some minor effect on the interaction. Most important, however, is the fact that the coffee did not simply merge with the ice: in transferring heat to the ice cube, the molecules in the coffee expended some of their internal kinetic energy, losing further heat in the process.

COOLING MACHINES. Even cooling machines, such as refrigerators and air conditioners, actually use heat, simply reversing the usual process by which particles are heated. The refrigerator pulls heat from its inner compartment—the area where food and other perishables are stored—and transfers it to the region outside. This is why the back of a refrigerator is warm.

Inside the refrigerator is an evaporator, into which heat from the refrigerated compartment flows. The evaporator contains a refrigerant—a gas, such as ammonia or Freon 12, that readily liquifies. This gas is released into a pipe from the evaporator at a low pressure, and as a result, it evaporates, a process that cools it. The pipe takes the refrigerant to the compressor, which pumps it into the condenser at a high pressure. Located at the back of the refrigerator, the condenser is a long series of pipes in which pressure turns the gas into liquid. As it moves through the condenser, the gas heats, and this heat is released into the air around the refrigerator.

An air conditioner works in a similar manner. Hot air from the room flows into the evaporator, and a compressor circulates refrigerant from the evaporator to a condenser. Behind the evaporator is a fan, which draws in hot air from the room, and another fan pushes heat from the condenser to the outside. As with a refrigerator, the back of an air conditioner is hot because it is moving heat from the area to be cooled.

Thus, cooling machines do not defy the principles of heat discussed above; nor do they defy the laws of thermodynamics that will be discussed at the conclusion of this essay. In accordance with the second law, in order to move heat in the reverse of its usual direction, external energy is required. Thus, a refrigerator takes in energy from a electric power supply (that is, the outlet it is plugged into), and extracts heat. Nonetheless, it manages to do so efficiently, removing two or three times as much heat from its inner compartment as the amount of energy required to run the refrigerator.

TRANSFERS OF HEAT

It is appropriate now to discuss how heat is transferred. One must remember, again, that in order for heat to be transferred from one point to another, there must be a difference of temperature between those two points. If an object or system has a uniform level of internal thermal energy—no matter how "hot" it may be in ordinary terms—no heat transfer is taking place.

Heat is transferred by one of three methods: conduction, which involves successive molecular collisions; convection, which requires the motion of hot fluid from one place to another; or radiation, which involves electromagnetic waves and requires no physical medium for the transfer.

CONDUCTION. Conduction takes place best in solids and particularly in metals, whose molecules are packed in relatively close proximity. Thus, when one end of an iron rod is heated, eventually the other end will acquire heat due to conduction. Molecules of liquid or nonmetallic solids vary in their ability to conduct heat, but gas—due to the loose attractions between its molecules—is a poor conductor.

When conduction takes place, it is as though a long line of people are standing shoulder to shoulder, passing a secret down the line. In this case, however, the "secret" is kinetic thermal energy. And just as the original phrasing of the secret will almost inevitably become garbled by the time it gets to the tenth or hundredth person, some energy is lost in the transfer from molecule to molecule. Thus, if one end of the iron rod is sitting in a fire and one end is surrounded by air at room temperature, it is unlikely that the end in the air will ever get as hot as the end in the fire.

Incidentally, the qualities that make metallic solids good conductors of heat also make them good conductors of electricity. In the first instance, kinetic energy is being passed from molecule to molecule, whereas in an electrical field, electrons—freed from the atoms of which they are normally a part—are able to move along the line of molecules. Because plastic is much less conductive than metal, an electrician will use a screwdriver with a plastic handle. Similarly, a metal pan typically has a handle of wood or plastic.

CONVECTION. There is a term, "convection oven," that is actually a redundancy: all ovens heat through convection, the principal means of transferring heat through a fluid. In physics, "fluid" refers both to liquids and gases—anything that tends to flow. Instead of simply moving heat, as in conduction, convection involves the movement of heated material—that is, fluid. When air is heated, it displaces cold (that is, unheated) air in its path, setting up a convection current.

Convection takes place naturally, as for instance when hot air rises from the land on a warm day. This heated air has a lower density than that of the less heated air in the atmosphere above it, and, therefore, is buoyant. As it rises, however, it loses energy and cools. This cooled air, now more dense than the air around it, sinks again, creating a repeating cycle.

The preceding example illustrates natural convection; the heat of an oven, on the other hand, is an example of forced convection—a situation in which some sort of pump or mechanism moves heated fluid. So, too, is the cooling work of a refrigerator, though the refrigerator moves heat in the opposite direction.

Forced convection can also take place within a natural system. The human heart is a pump, and blood carries excess heat generated by the body to the skin. The heat passes through the skin by means of conduction, and at the surface of the skin, it is removed from the body in a number of ways, primarily by the cooling evaporation of moisture—that is, perspiration.

RADIATION. If the Sun is hot—hot enough to severely burn the skin of a person who spends too much time exposed to its rays—then why is it cold in the upper atmosphere? After all, the upper atmosphere is closer to the Sun. And why is it colder still in the empty space above the atmosphere, which is still closer to the Sun? The reason is that in outer space there is no medium for convection, and in the upper atmosphere, where the air molecules are very far apart, there is hardly any medium. How, then, does heat come to the Earth from the Sun? By radiation, which is radically different from conduction or convection. The other two involve ordinary thermal energy, but radiation involves electromagnetic energy.

A great deal of "stuff" travels through the electromagnetic spectrum, discussed in another essay in this book: radio waves, microwaves for television and radar, infrared light, visible light, x rays, gamma rays. Though the relatively narrow band of visible-light wavelengths is the only part of the spectrum of which people are aware in everyday life, other parts—particularly the infrared and ultraviolet bands—are involved in the heat one feels from the Sun. (Ultraviolet rays, in fact, cause sunburns.)

Heat by means of radiation is not as "otherworldly" as it might seem: in fact, one does not have to point to the Sun for examples of it. Any time an object glows as a result of heat—as for example, in the case of firelight—that is an example of radiation. Some radiation is emitted in the form of visible light, but the heat component is in infrared rays. This also occurs in an incandescent light bulb. In an incandescent bulb, incidentally, much of the energy is lost to the heat of infrared rays, and the efficiency of a fluorescent bulb lies in the fact that it converts what would otherwise be heat into usable light.

THE LAWS OF THERMODYNAMICS

Having explored the behavior of heat, both at the molecular level and at levels more easily perceived by the senses, it is possible to discuss the laws of thermodynamics alluded to throughout this essay. These laws illustrate the relationships between heat and energy examined earlier, and show, for instance, why a refrigerator or air conditioner must have an external source of energy to move heat in a direction opposite to its normal flow.

BENJAMIN THOMPSON, COUNT RUMFORD. *(Illustration by H. Humphrey. UPI/Corbis-Bettmann. Reproduced by permission.)*

The story of how these laws came to be discovered is a saga unto itself, involving the contributions of numerous men in various places over a period of more than a century. In 1791, Swiss physicist Pierre Prevost (1751-1839) put forth his theory of exchanges, stating correctly that all bodies radiate heat. Hence, as noted earlier, there is no such thing as "cold": when one holds snow in one's hand, cold does not flow from the snow into the hand; rather, heat flows from the hand to the snow.

Seven years later, an American-British physicist named Benjamin Thompson, Count Rumford (1753) was boring a cannon with a blunt drill when he noticed that this action generated a great deal of heat. This led him to question the prevailing wisdom, which maintained that heat was a fluid form of matter; instead, Thompson began to suspect that heat must arise from some form of motion.

CARNOT'S ENGINE. The next major contribution came from the French physicist and engineer Sadi Carnot (1796-1832).

Though he published only one scientific work, *Reflections on the Motive Power of Fire* (1824), this treatise caused a great stir in the European scientific community. In it, Carnot made the first attempt at a scientific definition of work, describing it as "weight lifted through a height." Even more important was his proposal for a highly efficient steam engine.

A steam engine, like a modern-day internal combustion engine, is an example of a larger class of machine called heat engine. A heat engine absorbs heat at a high temperature, performs mechanical work, and, as a result, gives off heat a lower temperature. (The reason why that temperature must be lower is established in the second law of thermodynamics.)

For its era, the steam engine was what the computer is today: representing the cutting edge in technology, it was the central preoccupation of those interested in finding new ways to accomplish old tasks. Carnot, too, was fascinated by the steam engine, and was determined to help overcome its disgraceful inefficiency: in operation, a steam engine typically lost as much as 95% of its heat energy.

In his *Reflections,* Carnot proposed that the maximum efficiency of any heat engine was equal to $(T_H-T_L)/T_H$, where T_H is the highest operating temperature of the machine, and T_L the lowest. In order to maximize this value, T_L has to be absolute zero, which is impossible to reach, as was later illustrated by the third law of thermodynamics.

In attempting to devise a law for a perfectly efficient machine, Carnot inadvertently proved that such a machine is impossible. Yet his work influenced improvements in steam engine design, leading to levels of up to 80% efficiency. In addition, Carnot's studies influenced Kelvin—who actually coined the term "thermodynamics"—and others.

THE FIRST LAW OF THERMODYNAMICS. During the 1840s, Julius Robert Mayer (1814-1878), a German physicist, published several papers in which he expounded the principles known today as the conservation of energy and the first law of thermodynamics. As discussed earlier, the conservation of energy shows that within a system isolated from all outside factors, the total amount of energy remains the same, though transformations of energy from one form to another take place.

The first law of thermodynamics states this fact in a somewhat different manner. As with the other laws, there is no definitive phrasing; instead, there are various versions, all of which say the same thing. One way to express the law is as follows: Because the amount of energy in a system remains constant, it is impossible to perform work that results in an energy output greater than the energy input. For a heat engine, this means that the work output of the engine, combined with its change in internal energy, is equal to its heat input. Most heat engines, however, operate in a cycle, so there is no net change in internal energy.

Earlier, it was stated that a refrigerator extracts two or three times as much heat from its inner compartment as the amount of energy required to run it. On the surface, this seems to contradict the first law: isn't the refrigerator putting out more energy than it received? But the heat it extracts is only part of the picture, and not the most important part from the perspective of the first law.

A regular heat engine, such as a steam or internal-combustion engine, pulls heat from a high-temperature reservoir to a low-temperature reservoir, and, in the process, work is accomplished. Thus, the hot steam from the high-temperature reservoir makes possible the accomplishment of work, and when the energy is extracted from the steam, it condenses in the low-temperature reservoir as relatively cool water.

A refrigerator, on the other hand, reverses this process, taking heat from a low-temperature reservoir (the evaporator inside the cooling compartment) and pumping it to a high-temperature reservoir outside the refrigerator. Instead of producing a work output, as a steam engine does, it requires a work input—the energy supplied via the wall outlet. Of course, a refrigerator does produce an "output," by cooling the food inside, but the work it performs in doing so is equal to the energy supplied for that purpose.

THE SECOND LAW OF THERMODYNAMICS. Just a few years after Mayer's exposition of the first law, another German physicist, Rudolph Julius Emanuel Clausius (1822-1888) published an early version of the second law of thermodynamics. In an 1850 paper, Clausius stated that "Heat cannot, of itself, pass from a colder to a hotter body." He refined this 15 years later, introducing the concept of entropy—the tendency of natural systems toward breakdown, and specifically, the tendency for the energy in a system to be dissipated.

The second law of thermodynamics begins from the fact that the natural flow of heat is always from a high-temperature reservoir to a low-temperature reservoir. As a result, no engine can be constructed that simply takes heat from a source and performs an equivalent amount of work: some of the heat will always be lost. In other words, it is impossible to build a perfectly efficient engine.

Though its relation to the first law is obvious, inasmuch as it further defines the limitations of machine output, the second law of thermodynamics is not derived from the first. Elsewhere in this volume, the first law of thermodynamics—stated as the conservation of energy law—is discussed in depth, and, in that context, it is in fact necessary to explain how the behavior of machines in the real world does not contradict the conservation law.

Even though they mean the same thing, the first law of thermodynamics and the conservation of energy law are expressed in different ways. The first law of thermodynamics states that "the glass is half empty," whereas the conservation of energy law shows that "the glass is half full." The thermodynamics law emphasizes the bad news: that one can never get more energy out of a machine than the energy put into it. Thus, all hopes of a perpetual motion machine were dashed. The conservation of energy, on the other hand, stresses the good news: that energy is never lost.

In this context, the second law of thermodynamics delivers another dose of bad news: though it is true that energy is never lost, the energy available for work output will never be as great as the energy put into a system. A car engine, for instance, cannot transform all of its energy input into usable horsepower; some of the energy will be used up in the form of heat and sound. Though energy is conserved, usable energy is not.

Indeed, the concept of entropy goes far beyond machines as people normally understand them. Entropy explains why it is easier to break something than to build it—and why, for each person, the machine called the human body will inevitably break down and die, or cease to function, someday.

THE THIRD LAW OF THERMODYNAMICS. The subject of entropy leads directly to the third law of thermodynamics, formulated by German chemist Hermann Walter Nernst (1864-1941) in 1905. The third law states that at the temperature of absolute zero, entropy also approaches zero. From this statement, Nernst deduced that absolute zero is therefore impossible to reach.

All matter is in motion at the molecular level, which helps define the three major phases of matter found on Earth. At one extreme is a gas, whose molecules exert little attraction toward one another, and are therefore in constant motion at a high rate of speed. At the other end of the phase continuum (with liquids somewhere in the middle) are solids. Because they are close together, solid particles move very little, and instead of moving in relation to one another, they merely vibrate in place. But they do move.

Absolute zero, or 0K on the Kelvin scale of temperature, is the point at which all molecular motion stops entirely—or at least, it virtually stops. (In fact, absolute zero is defined as the temperature at which the motion of the average atom or molecule is zero.) As stated earlier, Carnot's engine achieves perfect efficiency if its lowest temperature is the same as absolute zero; but the second law of thermodynamics shows that a perfectly efficient machine is impossible. This means that absolute zero is an unreachable extreme, rather like matter exceeding the speed of light, also an impossibility.

This does not mean that scientists do not attempt to come as close as possible to absolute zero, and indeed they have come very close. In 1993, physicists at the Helsinki University of Technology Low Temperature Laboratory in Finland used a nuclear demagnetization device to achieve a temperature of $2.8 \cdot 10^{-10}$ K, or 0.00000000028K. This means that a fragment equal to only 28 parts in 100 billion separated this temperature from absolute zero—but it was still above 0K. Such extreme low-temperature research has a number of applications, most notably with superconductors, materials that exhibit virtually no resistance to electrical current at very low temperatures.

KEY TERMS

ABSOLUTE ZERO: The temperature, defined as 0K on the Kelvin scale, at which the motion of molecules in a solid virtually ceases. The third law of thermodynamics establishes the impossibility of actually reaching absolute zero.

BTU (BRITISH THERMAL UNIT): A measure of energy or heat in the British system, often used in reference to the capacity of an air conditioner. A Btu is equal to 778 foot-pounds, or 1,054 joules.

CALORIE: A measure of heat or energy in the SI or metric system, equal to the heat that must be added to or removed from 1 gram of water to change its temperature by 33.8°F (1°C). The dietary Calorie (capital C) with which most people are familiar is the same as the kilocalorie.

CONDUCTION: The transfer of heat by successive molecular collisions. Conduction is the principal means of heat transfer in solids, particularly metals.

CONSERVATION OF ENERGY: A law of physics which holds that within a system isolated from all other outside factors, the total amount of energy remains the same, though transformations of energy from one form to another take place. The first law of thermodynamics is the same as the conservation of energy.

CONSERVE: In physics, "to conserve" something means "to result in no net loss of" that particular component. It is possible that within a given system, the component may change form or position, but as long as the net value of the component remains the same, it has been conserved.

CONVECTION: The transfer of heat through the motion of hot fluid from one place to another. In physics, a "fluid" can be either a gas or a liquid, and convection is the principal means of heat transfer, for instance, in air and water.

ENERGY: The ability to accomplish work.

ENTROPY: The tendency of natural systems toward breakdown, and specifically, the tendency for the energy in a system to be dissipated. Entropy is closely related to the second law of thermodynamics.

FIRST LAW OF THERMODYNAMICS: A law which states the amount of energy in a system remains constant, and therefore it is impossible to perform work that results in an energy output greater than the energy input. This is the same as the conservation of energy.

FOOT-POUND: The principal unit of energy—and thus of heat—in the British or English system. The metric or SI unit is the joule. A foot-pound (ft • lb) is equal to 1.356 J.

HEAT: Internal thermal energy that flows from one body of matter to another. Heat is transferred by three methods conduction, convection, and radiation.

HEAT CAPACITY: The amount of heat that must be added to, or removed from, a unit of mass of a given substance to change its temperature by 33.8°F (1°C). Heat capacity is sometimes called specific heat capacity or specific heat. A kilocalorie is the heat capacity of 1 gram of water.

HEAT ENGINE: A machine that absorbs heat at a high temperature, performs mechanical work, and as a result gives off heat at a lower temperature.

KINETIC ENERGY: The energy that an object possesses by virtue of its motion.

KEY TERMS CONTINUED

JOULE: The principal unit of energy—and thus of heat—in the SI or metric system, corresponding to 1 newton-meter (N • m). A joule (J) is equal to 0.7376 foot-pounds.

KELVIN SCALE: Established by William Thomson, Lord Kelvin (1824-1907), the Kelvin scale measures temperature in relation to absolute zero, or 0K. (Units in the Kelvin system, known as Kelvins, do not include the word or symbol for degree.) The Kelvin and Celsius scales are directly related; hence Celsius temperatures can be converted to Kelvins by adding 273.15.

KILOCALORIE: A measure of heat or energy in the SI or metric system, equal to the heat that must be added to or removed from 1 kilogram of water to change its temperature by 33.8°F (1°C). As its name suggests, a kilocalorie is 1,000 calories. The dietary Calorie (capital C) with which most people are familiar is the same as the kilocalorie.

MECHANICAL ENERGY: The sum of potential energy and kinetic energy in a given system.

POTENTIAL ENERGY: The energy that an object possesses due to its position.

RADIATION: The transfer of heat by means of electromagnetic waves, which require no physical medium (e.g., water or air) for the transfer. Earth receives the Sun's heat by means of radiation.

SECOND LAW OF THERMODYNAMICS: A law of thermodynamics which states that no engine can be constructed that simply takes heat from a source and performs an equivalent amount of work. Some of the heat will always be lost, and therefore it is impossible to build a perfectly efficient engine. This is a result of the fact that the natural flow of heat is always from a high-temperature reservoir to a low-temperature reservoir—a fact expressed in the concept of entropy. The second law is sometimes referred to as "the law of entropy."

SYSTEM: In physics, the term "system" usually refers to any set of physical interactions isolated from the rest of the universe. Anything outside of the system, including all factors and forces irrelevant to a discussion of that system, is known as the environment.

TEMPERATURE: The direction of internal energy flow between bodies when heat is being transferred. Temperature measures the average molecular kinetic energy in transit between those bodies.

THERMAL ENERGY: Heat energy, a form of kinetic energy produced by the movement of atomic or molecular particles. The greater the movement of these particles, the greater the thermal energy.

THERMODYNAMICS: The study of the relationships between heat, work, and energy.

THIRD LAW OF THERMODYNAMICS: A law of thermodynamics which states that at the temperature of absolute zero, entropy also approaches zero. Zero entropy would contradict the second law of thermodynamics, meaning that absolute zero is therefore impossible to reach.

WORK: The exertion of force over a given distance to displace or move an object. Work is thus the product of force and distance exerted in the same direction.

WHERE TO LEARN MORE

Beiser, Arthur. *Physics,* 5th ed. Reading, MA: Addison-Wesley, 1991.

Brown, Warren. *Alternative Sources of Energy.* Introduction by Russell E. Train. New York: Chelsea House, 1994.

Encyclopedia of Thermodynamics (Web site). <http://therion.minpet.unibas.ch/minpet/groups/thermodict/> (April 12, 2001).

Entropy and the Second Law of Thermodynamics (Web site). <http://www.2ndlaw.com> (April 12, 2001).

Fleisher, Paul. *Matter and Energy: Principles of Matter and Thermodynamics.* Minneapolis, MN: Lerner Publications, 2002.

Macaulay, David. *The New Way Things Work.* Boston: Houghton Mifflin, 1998.

Moran, Jeffrey B. *How Do We Know the Laws of Thermodynamics?* New York: Rosen Publishing Group, 2001.

Santrey, Laurence. *Heat.* Illustrated by Lloyd Birmingham. Mahwah, N.J.: Troll Associates, 1985.

Suplee, Curt. *Everyday Science Explained.* Washington, D.C.: National Geographic Society, 1996.

"Temperature and Thermodynamics" PhysLINK.com (Web site). <http://www.physlink.com/ae_thermo.cfm> (April 12, 2001).

HEAT

CONCEPT

Heat is a form of energy—specifically, the energy that flows between two bodies because of differences in temperature. Therefore, the scientific definition of heat is different from, and more precise than, the everyday meaning. Physicists working in the area of thermodynamics study heat from a number of perspectives, including specific heat, or the amount of energy required to change the temperature of a substance, and calorimetry, the measurement of changes in heat as a result of physical or chemical changes. Thermodynamics helps us to understand such phenomena as the operation of engines and the gradual breakdown of complexity in physical systems—a phenomenon known as entropy.

HOW IT WORKS

Heat, Work, and Energy

Thermodynamics is the study of the relationships between heat, work, and energy. Work is the exertion of force over a given distance to displace or move an object, and is, thus, the product of force and distance exerted in the same direction. Energy, the ability to accomplish work, appears in numerous manifestations—including thermal energy, or the energy associated with heat.

Thermal and other types of energy, including electromagnetic, sound, chemical, and nuclear energy, can be described in terms of two extremes: kinetic energy, or the energy associated with movement, and potential energy, or the energy associated with position. If a spring is pulled back to its maximum point of tension, its potential energy is also at a maximum; once it is released and begins springing through the air to return to its original position, it begins gaining kinetic energy and losing potential energy.

All manifestations of energy appear in both kinetic and potential forms, somewhat like the way football teams are organized to play both offense or defense. Just as a football team takes an offensive role when it has the ball, and a defensive role when the other team has it, a physical system typically undergoes regular transformations between kinetic and potential energy, and may have more of one or the other, depending on what is taking place in the system.

What Heat Is and Is Not

Thermal energy is actually a form of kinetic energy generated by the movement of particles at the atomic or molecular level: the greater the movement of these particles, the greater the thermal energy. Heat is internal thermal energy that flows from one body of matter to another—or, more specifically, from a system at a higher temperature to one at a lower temperature. Thus, temperature, like heat, requires a scientific definition quite different from its common meaning: temperature measures the average molecular kinetic energy of a system, and governs the direction of internal energy flow between them.

Two systems at the same temperature are said to be in a state of thermal equilibrium. When this occurs, there is no exchange of heat. Though in common usage, "heat" is an expression of relative warmth or coldness, in physical terms, heat exists only in transfer between two systems. What people really mean by "heat" is the internal energy of a system—energy that is a property of that system rather than a property of transferred internal energy.

If you hold a snowball in your hand, as Vanna White and her son are doing in this picture, heat will move from your hand to the snowball. Your hand experiences this as a sensation of coldness. *(Reuters NewMedia Inc./Corbis. Reproduced by permission.)*

NO SUCH THING AS "COLD." Though the term "cold" has plenty of meaning in the everyday world, in physics terminology, it does not. Cold and heat are analogous to darkness and light: again, darkness means something in our daily experience, but in physical terms, darkness is simply the absence of light. To speak of cold or darkness as entities unto themselves is rather like saying, after spending 20 dollars, "I have 20 non-dollars in my pocket."

If you grasp a snowball in your hand, of course, your hand gets cold. The human mind perceives this as a transfer of cold from the snowball, but, in fact, exactly the opposite happens: heat moves from your hand to the snow, and if enough heat enters the snowball, it will melt. At the same time, the departure of heat from your hand results in a loss of internal energy near the surface of your hand, which you experience as a sensation of coldness.

Transfers of Heat

In holding the snowball, heat passes from the surface of the hand by one means, conduction, then passes through the snowball by another means, convection. In fact, there are three methods heat is transferred: conduction, involving successive molecular collisions and the transfer of heat between two bodies in contact; convection, which requires the motion of fluid from one place to another; or radiation, which takes place through electromagnetic waves and requires no physical medium, such as water or air, for the transfer.

CONDUCTION. Solids, particularly metals, whose molecules are packed relatively close together, are the best materials for conduction. Molecules of liquid or non-metallic solids vary in their ability to conduct heat, but gas is a poor conductor, because of the loose attractions between its molecules.

The qualities that make metallic solids good conductors of heat, as a matter of fact, also make them good conductors of electricity. In the conduction of heat, kinetic energy is passed from molecule to molecule, like a long line of people standing shoulder to shoulder, passing a secret. (And, just as the original phrasing of the secret becomes garbled, some kinetic energy is inevitably lost in the series of transfers.)

As for electrical conduction, which takes place in a field of electric potential, electrons are freed from their atoms; as a result, they are able to move along the line of molecules. Because plastic is much less conductive than metal, an electrician uses a screwdriver with a plastic handle; similarly, a metal cooking pan typically has a wooden or plastic handle.

CONVECTION. Wherever fluids are involved—and in physics, "fluid" refers both to liquids and gases—convection is a common form of heat transfer. Convection involves the movement of heated material—whether it is air, water, or some other fluid.

Convection is of two types: natural convection and forced convection, in which a pump or other mechanism moves the heated fluid. When heated air rises, this is an example of natural convection. Hot air has a lower density than that of the cooler air in the atmosphere above it, and, therefore, is buoyant; as it rises, however, it loses energy and cools. This cooled air, now denser than the air around it, sinks again, creating a repeating cycle that generates wind.

Examples of forced convection include some types of ovens and even a refrigerator or air conditioner. These two machines both move warm air from an interior to an exterior place. Thus, the refrigerator pulls hot air from the compartment and expels it to the surrounding room, while an air conditioner pulls heat from a building and releases it to the outside.

But forced convection does not necessarily involve humanmade machines: the human heart is a pump, and blood carries excess heat generated by the body to the skin. The heat passes through the skin by means of conduction, and at the surface of the skin, it is removed from the body in a number of ways, primarily by the cooling evaporation of perspiration.

RADIATION. Outer space, of course, is cold, yet the Sun's rays warm the Earth, an apparent paradox. Because there is no atmosphere in space, convection is impossible. In fact, heat from the Sun is not dependant on any fluid medium for its transfer: it comes to Earth by means of radiation. This is a form of heat transfer significantly different from the other two, because it involves electromagnetic energy, instead of ordinary thermal energy generated by the action of molecules. Heat from the Sun comes through a relatively narrow area of the light spectrum, including infrared, visible light, and ultraviolet rays.

Every form of matter emits electromagnetic waves, though their presence may not be readily perceived. Thus, when a metal rod is heated, it experiences conduction, but part of its heat is radiated, manifested by its glow—visible light. Even when the heat in an object is not visible, however, it may be radiating electromagnetic energy, for instance, in the form of infrared light. And, of course, different types of matter radiate better than others: in general, the better an object is at receiving radiation, the better it is at emitting it.

Measuring Heat

The measurement of temperature by degrees in the Fahrenheit or Celsius scales is a part of everyday life, but measurements of heat are not as familiar to the average person. Because heat is a form of energy, and energy is the ability to perform work, heat is, therefore, measured by the same units as work.

The principal unit of work or energy in the metric system (known within the scientific community as SI, or the SI system) is the joule. Abbreviated "J," a joule is equal to 1 newton-meter (N • m). The newton is the SI unit of force, and since work is equal to force multiplied by distance, measures of work can also be separated into these components. For instance, the British measure of work brings together a unit of distance, the foot, and a unit of force, the pound. A foot-pound (ft • lb) is equal to 1.356 J, and 1 joule is equal to 0.7376 ft • lb.

A refrigerator is a type of reverse heat engine that uses a compressor, like the one shown at the back of this refrigerator, to cool the refrigerator's interior. *(Ecoscene/Corbis. Reproduced by permission.)*

In the British system, Btu, or British thermal unit, is another measure of energy used for machines such as air conditioners. One Btu is equal to 778 ft • lb or 1,054 J. The kilocalorie in addition to the joule, is an important SI measure of heat. The amount of energy required to change the temperature of 1 gram of water by 1°C is called a calorie, and a kilocalorie is equal to 1,000 calories. Somewhat confusing is the fact that the dietary Calorie (capital C), with which most people are familiar, is not the same as a calorie (lowercase C)—rather, a dietary Calorie is the equivalent of a kilocalorie.

Real-Life Applications

Specific Heat

Specific heat is the amount of heat that must be added to, or removed from, a unit of mass for a given substance to change its temperature by 1°C. Thus, a kilocalorie, because it measures the amount of heat necessary to effect that change precisely for a kilogram of water, is identical to the specific heat for that particular substance in that particular unit of mass.

The higher the specific heat, the more resistant the substance is to changes in temperature. Many metals, in fact, have a low specific heat, making them easy to heat up and cool down. This contributes to the tendency of metals to expand when heated (a phenomenon also discussed in the Thermal Expansion essay), and, thus, to their malleability.

MEASURING AND CALCULATING SPECIFIC HEAT. The specific heat of any object is a function of its mass, its composition, and the desired change in temperature. The values of the initial and final temperature are not important—only the difference between them, which is the temperature change.

The components of specific heat are related to one another in the formula $Q = mc\delta T$. Here Q is the quantity of heat, measured in joules, which must be added. The mass of the object is designated by m, and the specific heat of the particular substance in question is represented with c. The Greek letter delta (δ) designates change, and δT stands for "change in temperature."

Specific heat is measured in units of J/kg • °C (joules per kilogram-degree Centigrade), though for the sake of convenience, this is usually rendered in terms of kilojoules (kJ), or 1,000 joules—that is, kJ/kg • °C. The specific heat of water is easily derived from the value of a kilocalorie: it is 4.185, the same number of joules required to equal a kilocalorie.

Calorimetry

The measurement of heat gain or loss as a result of physical or chemical change is called calorimetry (pronounced kal-IM-uh-tree). Like the word "calorie," the term is derived from a Latin root meaning "heat."

The foundations of calorimetry go back to the mid-nineteenth century, but the field owes much to scientists' work that took place over a period of about 75 years prior to that time. In 1780, French chemist Antoine Lavoisier (1743-1794) and French astronomer and mathematician Pierre Simon Laplace (1749-1827) had used a rudimentary ice calorimeter for measuring the heats in formations of compounds. Around the same time, Scottish chemist Joseph Black (1728-1799) became the first scientist to make a clear distinction between heat and temperature.

By the mid-1800s, a number of thinkers had come to the realization that—contrary to prevailing theories of the day—heat was a form of energy, not a type of material substance. Among these were American-British physicist Benjamin Thompson, Count Rumford (1753-1814) and English chemist James Joule (1818-1889)—for whom, of course, the joule is named.

Calorimetry as a scientific field of study actually had its beginnings with the work of French chemist Pierre-Eugene Marcelin Berthelot (1827-1907). During the mid-1860s, Berthelot became intrigued with the idea of measuring heat, and by 1880, he had constructed the first real calorimeter.

CALORIMETERS. Essential to calorimetry is the calorimeter, which can be any device for accurately measuring the temperature of a substance before and after a change occurs. A calorimeter can be as simple as a styrofoam cup. Its quality as an insulator, which makes styrofoam ideal for holding in the warmth of coffee and protecting the hand from scalding as well, also makes styrofoam an excellent material for calorimetric testing. With a styrofoam calorimeter, the temperature of the substance inside the cup is measured, a reaction is allowed to take place, and afterward, the temperature is measured a second time.

The most common type of calorimeter used is the bomb calorimeter, designed to measure the heat of combustion. Typically, a bomb calorimeter consists of a large container filled with water, into which is placed a smaller container, the combustion crucible. The crucible is made of metal, having thick walls with an opening through which oxygen can be introduced. In addition, the combustion crucible is designed to be connected to a source of electricity.

In conducting a calorimetric test using a bomb calorimeter, the substance or object to be studied is placed inside the combustion crucible and ignited. The resulting reaction usually occurs so quickly that it resembles the explosion of a bomb—hence, the name "bomb calorimeter." Once the "bomb" goes off, the resulting transfer of heat creates a temperature change in the water, which can be readily gauged with a thermometer.

To study heat changes at temperatures higher than the boiling point of water (212°F or 100°C), physicists use substances with higher boiling points. For experiments involving extremely large temperature ranges, an aneroid (without liquid) calorimeter may be used. In this case, the lining of the combustion crucible must be of a metal, such as copper, with a high coefficient or factor of thermal conductivity.

Heat Engines

The bomb calorimeter that Berthelot designed in 1880 measured the caloric value of fuels, and was applied to determining the thermal efficiency of a heat engine. A heat engine is a machine that absorbs heat at a high temperature, performs mechanical work, and as a result, gives off heat at a lower temperature.

The desire to create efficient heat engines spurred scientists to a greater understanding of thermodynamics, and this resulted in the laws of thermodynamics, discussed at the conclusion of this essay. Their efforts were intimately connected with one of the greatest heat engines ever created, a machine that literally powered the industrialized world during the nineteenth century: the steam engine.

HOW A STEAM ENGINE WORKS. Like all heat engines (except reverse heat engines such as the refrigerator, discussed below), a steam engine pulls heat from a high-temperature reservoir to a low-temperature reservoir, and in the process, work is accomplished. The hot steam from the high-temperature reservoir makes possible the accomplishment of work, and when the energy is extracted from the steam, the steam condenses in the low-temperature reservoir, becoming relatively cool water.

A steam engine is an external-combustion engine, as opposed to the internal-combustion engine that took its place at the forefront of industrial technology at the beginning of the twentieth century. Unlike an internal-combustion engine, a steam engine burns its fuel outside the engine. That fuel may be simply firewood, which is used to heat water and create steam. The thermal energy of the steam is then used to power a piston moving inside a cylinder, thus, converting thermal energy to mechanical energy for purposes such as moving a train.

EVOLUTION OF STEAM POWER. As with a number of advanced concepts in science and technology, the historical roots of the steam engine can be traced to the Greeks, who—just as they did with ideas such as the atom or the Sun-centered model of the universe—thought about it, but failed to develop it. The great inventor Hero of Alexandria (c. 65-125) actually created several steam-powered devices, but he perceived these as mere novelties, hardly worthy of scientific attention. Though Europeans adopted water power, as, for instance, in waterwheels, during the late ancient and medieval periods, further progress in steam power did not occur for some 1,500 years.

Following the work of French physicist Denis Papin (1647-1712), who invented the pressure cooker and conducted the first experiments with the use of steam to move a piston, English engineer Thomas Savery (c. 1650-1715) built the first steam engine. Savery had abandoned the use of the piston in his machine, but another English engineer, Thomas Newcomen (1663-1729), reintroduced the piston for his own steam-engine design.

Then in 1763, a young Scottish engineer named James Watt (1736-1819) was repairing a Newcomen engine and became convinced he could build a more efficient model. His steam engine, introduced in 1769, kept the heating and cooling processes separate, eliminating the need for the engine to pause in order to reheat. These and other innovations that followed—including the introduction of a high-pressure steam engine by English inventor Richard Trevithick (1771-1833)—transformed the world.

CARNOT PROVIDES THEORETICAL UNDERSTANDING. The men who developed the steam engine were mostly practical-minded figures who wanted only to build a better machine; they were not particularly concerned with the theoretical explanation for its workings. Then in 1824, a French physicist and engineer by the name of Sadi Carnot (1796-1832) published his sole work, the highly influential *Reflections on the Motive Power of Fire* (1824), in which he discussed heat engines scientifically.

In *Reflections,* Carnot offered the first definition of work in terms of physics, describing it as "weight lifted through a height." Analyzing Watt's steam engine, he also conducted groundbreaking studies in the nascent science of thermodynamics. Every heat engine, he explained, has a theoretical limit of efficiency related to the temperature difference in the engine: the greater the difference between the lowest and highest temperature, the more efficient the engine.

Carnot's work influenced the development of more efficient steam engines, and also had an impact on the studies of other physicists investigating the relationship between work, heat, and energy. Among these was William Thomson, Lord Kelvin (1824-1907). In addition to coining the term "thermodynamics," Kelvin developed the Kelvin scale of absolute temperature and established the value of absolute zero, equal to -273.15°C or -459.67°F.

According to Carnot's theory, maximum effectiveness was achieved by a machine that could reach absolute zero. However, later developments in the understanding of thermodynamics, as discussed below, proved that both maximum efficiency and absolute zero are impossible to attain.

REVERSE HEAT ENGINES. It is easy to understand that a steam engine is a heat engine: after all, it produces heat. But how is it that a refrigerator, an air conditioner, and other cooling machines are also heat engines? Moreover, given the fact that cold is the absence of heat and heat is energy, one might ask how a refrigerator or air conditioner can possibly use energy to produce cold, which is the same as the absence of energy. In fact, cooling machines simply reverse the usual process by which heat engines operate, and for this reason, they are called "reverse heat engines." Furthermore, they use energy to extract heat.

A steam engine takes heat from a high-temperature reservoir—the place where the water is turned into steam—and uses that energy to produce work. In the process, energy is lost and the heat moves to a low-temperature reservoir, where it condenses to form relatively cool water. A refrigerator, on the other hand, pulls heat from a low-temperature reservoir called the evaporator, into which flows heat from the refrigerated compartment—the place where food and other perishables are kept. The coolant from the evaporator take this heat to the condenser, a high-temperature reservoir at the back of the refrigerator, and in the process it becomes a gas. Heat is released into the surrounding air; this is why the back of a refrigerator is hot.

Instead of producing a work output, as a steam engine does, a refrigerator requires a work input—the energy supplied via the wall outlet. The principles of thermodynamics show that heat always flows from a high-temperature to a low-temperature reservoir, and reverse heat engines do not defy these laws. Rather, they require an external power source in order to effect the transfer of heat from a low-temperature reservoir, through the gases in the evaporator, to a high-temperature reservoir.

The Laws of Thermodynamics

THE FIRST LAW OF THERMODYNAMICS. There are three laws of thermodynamics, which provide parameters as to the operation of thermal systems in general, and heat engines in particular. The history behind the derivation of these laws is discussed in the essay on Thermodynamics; here, the laws themselves will be examined in brief form.

The physical law known as conservation of energy shows that within a system isolated from all outside factors, the total amount of energy remains the same, though transformations of energy from one form to another take place. The first law of thermodynamics states the same fact in a somewhat different manner.

According to the first law of thermodynamics, because the amount of energy in a system remains constant, it is impossible to perform work that results in an energy output greater than the energy input. Thus, it could be said that the conservation of energy law shows that "the glass is half full": energy is never lost. On the hand, the first law of thermodynamics shows that "the glass is half empty": no machine can ever produce more energy than was put into it. Hence, a perpetual motion machine is impossible, because in order to keep a machine running continually, there must be a continual input of energy.

THE SECOND LAW OF THERMODYNAMICS. The second law of ther-

KEY TERMS

ABSOLUTE ZERO: The temperature, defined as 0K on the Kelvin scale, at which the motion of molecules in a solid virtually ceases. The third law of thermodynamics establishes the impossibility of actually reaching absolute zero.

BTU (BRITISH THERMAL UNIT): A measure of energy or heat in the British system, often used in reference to the capacity of an air conditioner. A Btu is equal to 778 foot-pounds, or 1,054 joules.

CALORIE: A measure of heat or energy in the SI or metric system, equal to the heat that must be added to or removed from 1 gram of water to change its temperature by 1°C. The dietary Calorie (capital C) with which most people are familiar is the same as the kilocalorie.

CALORIMETRY: The measurement of heat gain or loss as a result of physical or chemical change.

CONDUCTION: The transfer of heat by successive molecular collisions. Conduction is the principal means of heat transfer in solids, particularly metals.

CONSERVATION OF ENERGY: A law of physics stating that within a system isolated from all other outside factors, the total amount of energy remains the same, though transformations of energy from one form to another take place. The first law of thermodynamics is the same as the conservation of energy.

CONVECTION: The transfer of heat through the motion of hot fluid from one place to another. In physics, a "fluid" can be either a gas or a liquid, and convection is the principal means of heat transfer, for instance, in air and water.

ENERGY: The ability to accomplish work.

ENTROPY: The tendency of natural systems toward breakdown, and specifically, the tendency for the energy in a system to be dissipated. Entropy is closely related to the second law of thermodynamics.

FIRST LAW OF THERMODYNAMICS: A law stating that the amount of energy in a system remains constant, and therefore, it is impossible to perform work that results in an energy, output greater than the energy input. This is the same as the conservation of energy.

FOOT-POUND: The principal unit of energy—and, thus, of heat—in the British or English system. The metric or SI unit is the joule. A foot-pound (ft • lb) is equal to 1.356 J.

HEAT: Internal thermal energy that flows from one body of matter to another. Heat is transferred by three methods conduction, convection, and radiation.

HEAT ENGINE: A machine that absorbs heat at a high temperature, performs mechanical work, and, as a result, gives off heat at a lower temperature.

JOULE: The principal unit of energy—and, thus, of heat—in the SI or metric system, corresponding to 1 newton-meter (N • m). A joule (J) is equal to 0.7376 foot-pounds.

KEY TERMS CONTINUED

KELVIN SCALE: Established by William Thomson, Lord Kelvin (1824-1907), the Kelvin scale measures temperature in relation to absolute zero, or 0K. (Units in the Kelvin system, known as Kelvins, do not include the word or symbol for degree.) The Kelvin and Celsius scales are directly related; hence, Celsius temperatures can be converted to Kelvins by adding 273.15.

KILOCALORIE: A measure of heat or energy in the SI or metric system, equal to the heat that must be added to or removed from 1 kilogram of water to change its temperature by 1°C. As its name suggests, a kilocalorie is 1,000 calories. The dietary Calorie (capital C) with which most people are familiar, is the same as the kilocalorie.

KINETIC ENERGY: The energy that an object possesses by virtue of its motion.

POTENTIAL ENERGY: The energy that an object possesses due to its position.

RADIATION: The transfer of heat by means of electromagnetic waves, which require no physical medium (for example, water or air) for the transfer. Earth receives the Sun's heat by means of radiation.

SECOND LAW OF THERMODYNAMICS: A law of thermodynamics stating that no engine can be constructed that simply takes heat from a source and performs an equivalent amount of work. Some of the heat will always be lost, and, therefore, it is impossible to build a perfectly efficient engine. This is a result of the fact that the natural flow of heat is always from a high-temperature reservoir to a low-temperature reservoir—a fact expressed in the concept of entropy. The second law is sometimes referred to as "the law of entropy."

modynamics begins from the fact that the natural flow of heat is always from a high-temperature to a low-temperature reservoir. As a result, no engine can be constructed that simply takes heat from a source and performs an equivalent amount of work: some of the heat will always be lost. In other words, it is impossible to build a perfectly efficient engine.

In effect, the second law of thermodynamics compounds the "bad news" delivered by the first law with some even worse news: though it is true that energy is never lost, the energy available for work output will never be as great as the energy put into a system. Linked to the second law is the concept of entropy, the tendency of natural systems toward breakdown, and specifically, the tendency for the energy in a system to be dissipated. "Dissipated" in this context means that the high- and low-temperature reservoirs approach equal temperatures, and as this occurs, entropy increases.

THE THIRD LAW OF THERMODYNAMICS. Entropy also plays a part in the third law of thermodynamics, which states that at the temperature of absolute zero, entropy also approaches zero. This might seem to counteract the "worse news" of the second law, but in fact, what the third law shows is that absolute zero is impossible to reach.

As stated earlier, Carnot's engine would achieve perfect efficiency if its lowest temperature were the same as absolute zero; but the second law of thermodynamics shows that a perfectly efficient machine is impossible. Relativity theory (which first appeared in 1905, the same year as the third law of thermodynamics) showed that matter can never exceed the speed of light. In the same way, the collective effect of the second and third laws is to prove that absolute

KEY TERMS CONTINUED

SPECIFIC HEAT: The amount of heat that must be added to, or removed from, a unit of mass of a given substance to change its temperature by 1°C. A kilocalorie is the specific heat of 1 gram of water.

SYSTEM: In physics, the term "system" usually refers to any set of physical interactions isolated from the rest of the universe. Anything outside of the system, including all factors and forces irrelevant to a discussion of that system, is known as the environment.

TEMPERATURE: The direction of internal energy flow between two systems when heat is being transferred. Temperature measures the average molecular kinetic energy in transit between those systems.

THERMAL ENERGY: Heat energy, a form of kinetic energy produced by the movement of atomic or molecular particles. The greater the movement of these particles, the greater the thermal energy.

THERMAL EQUILIBRIUM: The state that exists when two systems have the same temperature. As a result, there is no exchange of heat between them.

THERMODYNAMICS: The study of the relationships between heat, work, and energy.

THIRD LAW OF THERMODYNAMICS: A law of thermodynamics which states that at the temperature of absolute zero, entropy also approaches zero. Zero entropy would contradict the second law of thermodynamics, meaning that absolute zero is, therefore, impossible to reach.

WORK: The exertion of force over a given distance to displace or move an object. Work is, thus, the product of force and distance exerted in the same direction.

zero—the temperature at which molecular motion in all forms of matter theoretically ceases—can never be reached.

WHERE TO LEARN MORE

Beiser, Arthur. *Physics,* 5th ed. Reading, MA: Addison-Wesley, 1991.

Bonnet, Robert L and Dan Keen. *Science Fair Projects: Physics.* Illustrated by Frances Zweifel. New York: Sterling, 1999.

Encyclopedia of Thermodynamics (Web site). <http://therion.minpet.unibas.ch/minpet/groups/thermodict/> (April 12, 2001).

Friedhoffer, Robert. *Physics Lab in the Home.* Illustrated by Joe Hosking. New York: Franklin Watts, 1997.

Manning, Mick and Brita Granström. *Science School.* New York: Kingfisher, 1998.

Macaulay, David. *The New Way Things Work.* Boston: Houghton Mifflin, 1998.

Moran, Jeffrey B. *How Do We Know the Laws of Thermodynamics?* New York: Rosen Publishing Group, 2001.

Santrey, Laurence. *Heat.* Illustrated by Lloyd Birmingham. Mahwah, NJ: Troll Associates, 1985.

Suplee, Curt. *Everyday Science Explained.* Washington, D.C.: National Geographic Society, 1996.

"Temperature and Thermodynamics" PhysLINK.com (Web site). <http://www.physlink.com/ae_thermo.cfm> (April 12, 2001).

TEMPERATURE

CONCEPT

Temperature is one of those aspects of the everyday world that seems rather abstract when viewed from the standpoint of physics. In scientific terms, it is not simply a measure of hot and cold, but an indicator of molecular motion and energy flow. Thermometers measure temperature by a number of means, including the expansion that takes place in a medium such as mercury or alcohol. These measuring devices are gauged in several different ways, with scales based on the freezing and boiling points of water—as well as, in the case of the absolute temperature scale, the point at which all molecular motion virtually ceases.

HOW IT WORKS

Heat

Energy appears in many forms, including thermal energy, or the energy associated with heat. Heat is internal thermal energy that flows from one body of matter to another—or, more specifically, from a system at a higher temperature to one at a lower temperature.

Two systems at the same temperature are said to be in a state of thermal equilibrium. When this occurs, there is no exchange of heat. Though people ordinarily speak of "heat" as an expression of relative warmth or coldness, in physical terms, heat only exists in transfer between two systems. It is never something inherently part of a system; thus, unless there is a transfer of internal energy, there is no heat, scientifically speaking.

HEAT: ENERGY IN TRANSIT. Thus, heat cannot be said to exist unless there is one system in contact with another system of differing temperature. This can be illustrated by way of the old philosophical question: "If a tree falls in the woods when there is no one to hear it, does it make a sound?" From a physicist's point of view, of course, sound waves are emitted whether or not there is an ear to receive their vibrations; but, consider this same scenario in terms of heat. First, replace the falling tree with a hypothetical object possessing a certain amount of internal energy; then replace sound waves with heat. In this case, if this object is not in contact with something else that has a different temperature, it "does not make a sound"—in other words, it transfers no internal energy, and, thus, there is no heat from the standpoint of physics.

This could even be true of two incredibly "hot" objects placed next to one another inside a vacuum—an area devoid of matter, including air. If both have the same temperature, there is no heat, only two objects with high levels of internal energy. Note that a vacuum was specified: assuming there was air around them, and that the air was of a lower temperature, both objects would then be transferring heat to the air.

RELATIVE MOTION BETWEEN MOLECULES. If heat is internal thermal energy in transfer, from whence does this energy originate? From the movement of molecules. Every type of matter is composed of molecules, and those molecules are in motion relative to one another. The greater the amount of relative motion between molecules, the greater the kinetic energy, or the energy of movement, which is manifested as thermal energy. Thus, "heat"—to use the everyday term for what physicists describe as thermal energy—is really nothing more than the result of relative molecular motion. Thus, thermal energy is sometimes identified as molecular translational energy.

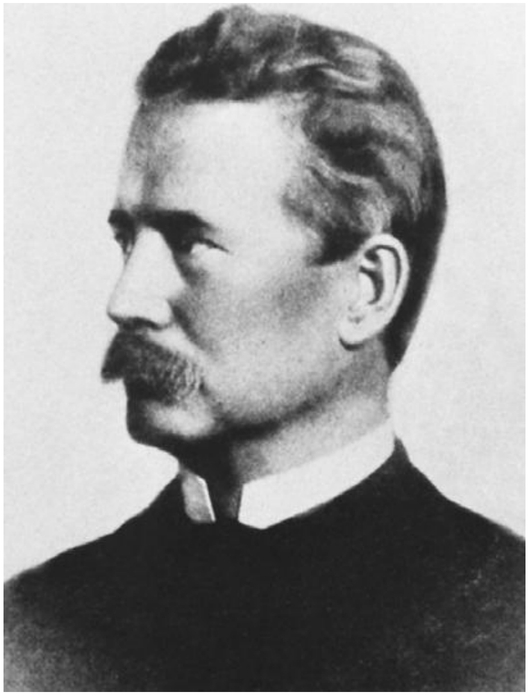
Note that the molecules are in relative motion, meaning that if one were "standing" on a molecule, one would see the other molecules moving. This is not the same as movement on the part of a large object composed of molecules; in this case, molecules themselves are not directly involved in relative motion.

Put another way, the movement of Earth through space is an entirely different type of movement from the relative motion of objects on Earth—people, animals, natural forms such as clouds, manmade forms of transportation, and so forth. In this example, Earth is analogous to a "large" item of matter, such as a baseball, a stream of water, or a cloud of gas.

The smaller objects on Earth are analogous to molecules, and, in both cases, the motion of the larger object has little direct impact on the motion of smaller objects. Hence, as discussed in the Frame of Reference essay, it is impossible to perceive with one's senses the fact that Earth is actually hurling through space at incredible speeds.

MOLECULAR MOTION AND PHASES OF MATTER. The relative motion of molecules determines phase of matter—that is, whether something is a solid, liquid, or gas. When molecules move quickly in relation to one another, they exert a small electromagnetic attraction toward one another, and the larger material of which they are a part is called a gas. A liquid, on the other hand, is a type of matter in which molecules move at moderate speeds in relation to one another, and therefore exert a moderate intermolecular attraction.

The kinetic theory of gases relates molecular motion to energy in gaseous substances. It does not work as well in relation to liquids and solids; nonetheless, it is safe to say that—generally speaking—a gas has more energy than a liquid, and a liquid more energy than a solid. In a solid, the molecules undergo very little relative motion: instead of bumping into each other, like gas molecules and (to a lesser extent) liquid molecules, solid molecules merely vibrate in place.



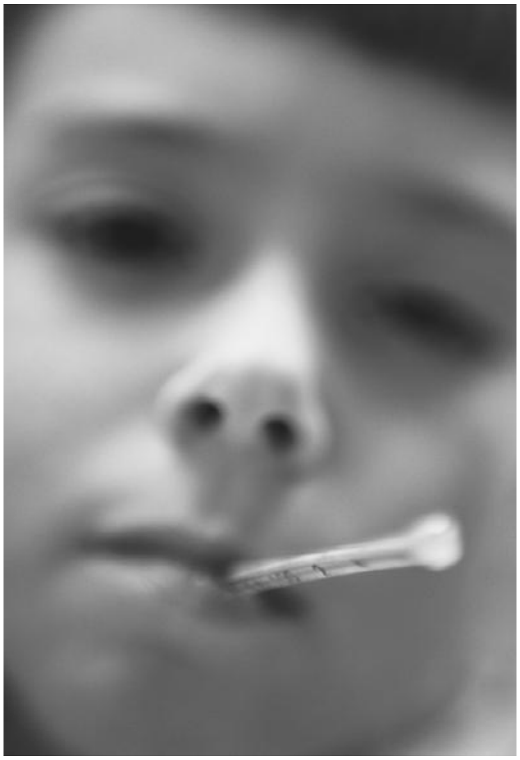
William Thomson, (better known as Lord Kelvin) established what is now known as the Kelvin scale.

Understanding Temperature

As with heat, temperature requires a scientific definition quite different from its common meaning. Temperature may be defined as a measure of the average molecular translational energy in a system—that is, in any material body.

Because it is an average, the mass or other characteristics of the body do not matter. A large quantity of one substance, because it has more molecules, possesses more thermal energy than a smaller quantity of that same substance. Since it has more thermal energy, it transfers more heat to any body or system with which it is in contact. Yet, assuming that the substance is exactly the same, the temperature, as a measure of average energy, will be the same as well.

Temperature determines the direction of internal energy flow between two systems when heat is being transferred. This can be illustrated through an experience familiar to everyone: having one's temperature taken with a thermometer. If one has a fever, one's mouth will be warmer than the thermometer, and therefore heat will be transferred to the thermometer from the mouth until the two objects have the same temperature.



A thermometer works by measuring the level of thermal expansion experienced by a material within the thermometer. Mercury has been a common thermometer material since the 1700s. *(Photograph by Michael Prince/Corbis. Reproduced by permission.)*

At that point of thermal equilibrium, a temperature reading can be taken from the thermometer.

TEMPERATURE AND HEAT FLOW. The principles of thermodynamics—the study of the relationships between heat, work, and energy, show that heat always flows from an area of higher temperature to an area of lower temperature. The opposite simply cannot happen, because coldness, though it is very real in terms of sensory experience, is not an independent phenomenon. There is not, strictly speaking, such a thing as "cold"—only the absence of heat, which produces the sensation of coldness.

One might pour a kettle of boiling water into a cold bathtub to heat it up; or put an ice cube in a hot cup of coffee "to cool it down." These seem like two very different events, but from the standpoint of thermodynamics, they are exactly the same. In both cases, a body of high temperature is placed in contact with a body of low temperature, and in both cases, heat passes from the high-temperature body to the low-temperature one.

The boiling water warms the tub of cool water, and due to the high ratio of cool water to boiling water in the bathtub, the boiling water expends all its energy raising the temperature in the bathtub as a whole. The greater the ratio of very hot water to cool water, on the other hand, the warmer the bathtub will be in the end. But even after the bath water is heated, it will continue to lose heat, assuming the air in the room is not warmer than the water in the tub. If the water in the tub is warmer than the air, it immediately begins transferring thermal energy to the low-temperature air until their temperatures are equalized.

As for the coffee and the ice cube, what happens is quite different from, indeed, opposite to, the common understanding of the process. In other words, the ice does not "cool down" the coffee: the coffee warms up the ice and presumably melts it. Once again, however, it expends at least some of its thermal energy in doing so, and as a result, the coffee becomes cooler than it was.

If the coffee is placed inside a freezer, there is a large temperature difference between it and the surrounding environment—so much so that if it is left for hours, the once-hot coffee will freeze. But again, the freezer does not cool down the coffee; the molecules in the coffee respond to the temperature difference by working to warm up the freezer. In this case, they have to "work overtime," and since the freezer has a constant supply of electrical energy, the heated molecules of the coffee continue to expend themselves in a futile effort to warm the freezer. Eventually, the coffee loses so much energy that it is frozen solid; meanwhile, the heat from the coffee has been transferred outside the freezer to the atmosphere in the surrounding room.

THERMAL EXPANSION AND EQUILIBRIUM. Temperature is related to the concept of thermal equilibrium, and has an effect on thermal expansion. As discussed below, as well as within the context of thermal expansion, a thermometer provides a gauge of temperature by measuring the level of thermal expansion experienced by a material (for example, mercury) within the thermometer.

In the examples used earlier—the thermometer in the mouth, the hot water in the cool bathtub, and the ice cube in the cup of coffee—the systems in question eventually reach thermal equilibrium. This is rather like averaging their temperatures, though, in fact, the equation involved is more complicated than a simple arithmetic average.

In the case of an ordinary mercury thermometer, the need to achieve thermal equilibrium explains why one cannot get an instantaneous temperature reading: first, the mouth transfers heat to the thermometer, and once both mouth and thermometer reach the same temperature, they are in thermal equilibrium. At that point, it is possible to gauge the temperature of the mouth by reading the thermometer.

REAL-LIFE APPLICATIONS

Development of the Thermometer

A thermometer can be defined scientifically as a device that gauges temperature by measuring a temperature-dependent property, such as the expansion of a liquid in a sealed tube. As with many aspects of scientific or technological knowledge, the idea of the thermometer appeared in ancient times, but was never developed. Again, like so many other intellectual phenomena, it lay dormant during the medieval period, only to be resurrected at the beginning of the modern era.

The Greco-Roman physician Galen (c. 129-216) was among the first thinkers to envision a scale for measuring temperature. Of course, what he conceived of as "temperature" was closer to the everyday meaning of that term, not its more precise scientific definition: the ideas of molecular motion, heat, and temperature discussed in this essay emerged only in the period beginning about 1750. In any case, Galen proposed that equal amounts of boiling water and ice be combined to establish a "neutral" temperature, with four units of warmth above it and four degrees of cold below.

THE THERMOSCOPE. The great physicist Galileo Galilei (1564-1642) is sometimes credited with creating the first practical temperature measuring device, called a thermoscope. Certainly Galileo—whether or not he was the first—did build a thermoscope, which consisted of a long glass tube planted in a container of liquid. Prior to inserting the tube into the liquid—which was usually colored water, though Galileo's thermoscope used wine—as much air as possible was removed from the tube. This created a vacuum, and as a result of pressure differences between the liquid and the interior of the thermoscope tube, some of the liquid went into the tube.

But the liquid was not the thermometric medium—that is, the substance whose temperature-dependent property changes the thermoscope measured. (Mercury, for instance, is the thermometric medium in most thermometers today.) Instead, the air was the medium whose changes the thermoscope measured: when it was warm, the air expanded, pushing down on the liquid; and when the air cooled, it contracted, allowing the liquid to rise.

It is interesting to note the similarity in design between the thermoscope and the barometer, a device for measuring atmospheric pressure invented by Italian physicist Evangelista Torricelli (1608-1647) around the same time. Neither were sealed, but by the mid-seventeenth century, scientists had begun using sealed tubes containing liquid instead of air. These were the first true thermometers.

EARLY THERMOMETERS. Ferdinand II, Grand Duke of Tuscany (1610-1670), is credited with developing the first thermometer in 1641. Ferdinand's thermometer used alcohol sealed in glass, which was marked with a temperature scale containing 50 units. It did not, however, designate a value for zero.

English physicist Robert Hooke (1635-1703) created a thermometer using alcohol dyed red. Hooke's scale was divided into units equal to about 1/500 of the volume of the thermometric medium, and for the zero point, he chose the temperature at which water freezes. Thus, Hooke established a standard still used today; likewise, his thermometer itself set a standard. Built in 1664, it remained in use by the Royal Society—the foremost organization for the advancement of science in England during the early modern period—until 1709.

Olaus Roemer (1644-1710), a Danish astronomer, introduced another important standard. In 1702, he built a thermometer based not on one but two fixed points, which he designated as the temperature of snow or crushed ice, and the boiling point of water. As with Hooke's use of the freezing point, Roemer's idea of the freezing and boiling points of water as the two parameters for temperature measurements has remained in use ever since.

Temperature Scales

FAHRENHEIT. Not only did he develop the Fahrenheit scale, oldest of the temperature scales still used in Western nations today, but German physicist Daniel Fahrenheit (1686-1736) also built the first thermometer to contain mercury as a thermometric medium. Alcohol has a low boiling point, whereas mercury remains fluid at a wide range of temperatures. In addition, it expands and contracts at a very constant rate, and tends not to stick to glass. Furthermore, its silvery color makes a mercury thermometer easy to read.

Fahrenheit also conceived the idea of using "degrees" to measure temperature in his thermometer, which he introduced in 1714. It is no mistake that the same word refers to portions of a circle, or that exactly 180 degrees—half the number in a circle—separate the freezing and boiling points for water on Fahrenheit's thermometer. Ancient astronomers attempting to denote movement in the skies used a circle with 360 degrees as a close approximation of the ratio between days and years. The number 360 is also useful for computations, because it has a large quantity of divisors, as does 180—a total of 16 whole-number divisors other than 1 and itself.

Though it might seem obvious that 0 should denote the freezing point and 180 the boiling point on Fahrenheit's scale, such an idea was far from obvious in the early eighteenth century. Fahrenheit considered the idea not only of a 0-to-180 scale, but also of a 180-to-360 scale. In the end, he chose neither—or rather, he chose not to equate the freezing point of water with zero on his scale. For zero, he chose the coldest possible temperature he could create in his laboratory, using what he described as "a mixture of sal ammoniac or sea salt, ice, and water." Salt lowers the melting point of ice (which is why it is used in the northern United States to melt snow and ice from the streets on cold winter days), and, thus, the mixture of salt and ice produced an extremely cold liquid water whose temperature he equated to zero.

With Fahrenheit's scale, the ordinary freezing point of water was established at 32°, and the boiling point exactly 180° above it, at 212°. Just a few years after he introduced his scale, in 1730, a French naturalist and physicist named Rene Antoine Ferchault de Reaumur (1683-1757) presented a scale for which 0° represented the freezing point of water and 80° the boiling point. Although the Reaumur scale never caught on to the same extent as Fahrenheit's, it did include one valuable addition: the specification that temperature values be determined at standard sea-level atmospheric pressure.

CELSIUS. With its 32-degree freezing point and its 212-degree boiling point, the Fahrenheit system is rather ungainly, lacking the neat orderliness of a decimal or base-10 scale. The latter quality became particularly important when, 10 years after the French Revolution of 1789, France adopted the metric system for measuring length, mass, and other physical phenomena. The metric system eventually spread to virtually the entire world, with the exception of English-speaking countries, where the more cumbersome British system still prevails. But even in the United States and Great Britain, scientists use the metric system. The metric temperature measure is the Celsius scale, created in 1742 by Swedish astronomer Anders Celsius (1701-1744).

Like Fahrenheit, Celsius chose the freezing and boiling points of water as his two reference points, but he determined to set them 100, rather than 180, degrees apart. Interestingly, he planned to equate 0° with the boiling point, and 100° with the freezing point—proving that even the most apparently obvious aspects of a temperature scale were once open to question. Only in 1750 did fellow Swedish physicist Martin Strömer change the orientation of the Celsius scale.

Celsius's scale was based not simply on the boiling and freezing points of water, but, specifically, those points at normal sea-level atmospheric pressure. The latter, itself a unit of measure known as an atmosphere (atm), is equal to 14.7 lb/in^2, or 101,325 pascals in the metric system. A Celsius degree is equal to 1/100 of the difference between the freezing and boiling temperatures of water at 1 atm.

The Celsius scale is sometimes called the centigrade scale, because it is divided into 100 degrees, *cent* being a Latin root meaning "hundred." By international convention, its values were refined in 1948, when the scale was redefined in terms of temperature change for an ideal gas, as well as the triple point of water. (Triple point is the temperature and pressure at which a substance is at once a solid, liquid, and vapor.) As a result of these refinements, the boiling point of water on the Celsius scale is actually 99.975°. This represents a difference equal to about 1 part in 4,000—hardly significant in daily life, though a significant change from the standpoint of the precise measurements made in scientific laboratories.

KELVIN. In about 1787, French physicist and chemist J. A. C. Charles (1746-1823) made an interesting discovery: that at 0°C, the volume of gas at constant pressure drops by 1/273 for every Celsius degree drop in temperature. This seemed to suggest that the gas would simply disappear if cooled to -273°C, which, of course, made no sense. In any case, the gas would most likely become first a liquid, and then a solid, long before it reached that temperature.

The man who solved the quandary raised by Charles's discovery was born a year after Charles—who also formulated Charles's law—died. He was William Thomson, Lord Kelvin (1824-1907), and in 1848, he put forward the suggestion that it was molecular translational energy, and not volume, that would become zero at -273°C. He went on to establish what came to be known as the Kelvin scale.

Sometimes known as the absolute temperature scale, the Kelvin scale is based not on the freezing point of water, but on absolute zero—the temperature at which molecular motion comes to a virtual stop. This is -273.15°C (-459.67°F), which in the Kelvin scale is designated as 0K. (Kelvin measures do not use the term or symbol for "degree.")

Though scientists normally use metric or SI measures, they prefer the Kelvin scale to Celsius, because the absolute temperature scale is directly related to average molecular translational energy. Thus, if the Kelvin temperature of an object is doubled, this means that its average molecular translational energy has doubled as well. The same cannot be said if the temperature were doubled from, say, 10°C to 20°C, or from 40°C to 80°F, since neither the Celsius nor the Fahrenheit scale is based on absolute zero.

CONVERSIONS. The Kelvin scale is, however, closely related to the Celsius scale, in that a difference of 1 degree measures the same amount of temperature in both. Therefore, Celsius temperatures can be converted to Kelvins by adding 273.15. There is also an absolute temperature scale that uses Fahrenheit degrees. This is the Rankine scale, created by Scottish engineer William Rankine (1820-1872), but it is seldom used today: scientists and others who desire absolute temperature measures prefer the precision and simplicity of the Celsius-based Kelvin scale.

Conversion between Celsius and Fahrenheit figures is a bit more challenging. To convert a temperature from Celsius to Fahrenheit, multiply by 9/5 and add 32. It is important to perform the steps in that order, because reversing them will produce a wrong answer. Thus, 100°C multiplied by 9/5 or 1.8 equals 180, which, when added to 32 equals 212°F. Obviously, this is correct, since 100°C and 212°F each represent the boiling point of water. But, if one adds 32 to 100°, then multiplies it by 9/5, the result is 237.6°F—an incorrect answer.

For converting Fahrenheit temperatures to Celsius, there are also two steps, involving multiplication and subtraction, but the order is reversed. Here, the subtraction step is performed before the multiplication step: thus, 32 is subtracted from the Fahrenheit temperature, then the result is multiplied by 5/9. Beginning with 212°F, if 32 is subtracted, this equals 180. Multiplied by 5/9, the result is 100°C—the correct answer.

One reason the conversion formulae use fractions instead of decimal fractions (what most people simply call "decimals") is that 5/9 is a repeating decimal fraction (0.55555....) Furthermore, the symmetry of 5/9 and 9/5 makes memorization easy. One way to remember the formula is that *F*ahrenheit is multiplied by a *f*raction—since 5/9 is a real fraction, whereas 9/5 is actually a whole number plus a fraction.

Thermometers

As discussed earlier, with regard to the early history of the thermometer, it is important that the glass tube be kept sealed; otherwise, atmospheric pressure contributes to inaccurate readings, because it influences the movement of the thermometric medium. Also important is the choice of the thermometric medium itself.

Water quickly proved unreliable, due to its unusual properties: it does not expand uniformly with a rise in temperature, or contract uniformly with a lowered temperature. Rather, it

KEY TERMS

ABSOLUTE ZERO: The temperature, defined as 0K on the Kelvin scale, at which the motion of molecules in a solid virtually ceases.

CELSIUS SCALE: A scale of temperature, sometimes known as the centigrade scale, created in 1742 by Swedish astronomer Anders Celsius (1701-1744). The Celsius scale establishes the freezing and boiling points of water at 0° and 100°, respectively. To convert a temperature from the Celsius to the Fahrenheit scale, multiply by 9/5 and add 32. The Celsius scale is part of the metric system used by most non-English speaking countries today. Though the worldwide scientific community uses the metric or SI system for most measurements, scientists prefer the related Kelvin scale.

FAHRENHEIT SCALE: The oldest of the temperature scales still used in Western nations today, created in 1714 by German physicist Daniel Fahrenheit (1686-1736). The Fahrenheit scale establishes the freezing and boiling points of water at 32° and 212° respectively. To convert a temperature from the Fahrenheit to the Celsius scale, subtract 32 and multiply by 5/9. Most English-speaking countries use the Fahrenheit scale.

HEAT: Internal thermal energy that flows from one body of matter to another.

KELVIN SCALE: Established by William Thomson, Lord Kelvin (1824-1907), the Kelvin scale measures temperature in relation to absolute zero, or 0K. (Units in the Kelvin system, known as Kelvins, do not include the word or symbol for degree.) The Kelvin and Celsius scales are directly related; hence, Celsius temperatures can be converted to Kelvins by adding 273.15. The Kelvin scale is used almost exclusively by scientists.

KINETIC ENERGY: The energy that an object possesses by virtue of its motion.

reaches its maximum density at 39.2°F (4°C), and is less dense both above and below that temperature. Therefore, alcohol, which responds in a much more uniform fashion to changes in temperature, took its place.

MERCURY THERMOMETERS. Alcohol is still used in thermometers today, but the preferred thermometric medium is mercury. As noted earlier, its advantages include a much higher boiling point, a tendency not to stick to glass, and a silvery color that makes its levels easy to gauge visually. Like alcohol, mercury expands at a uniform rate with an increase in temperature: hence, the higher the temperature, the higher the mercury stands in the thermometer.

In a typical mercury thermometer, mercury is placed in a long, narrow sealed tube called a capillary. The capillary is inscribed with figures for a calibrated scale, usually in such a way as to allow easy conversions between Fahrenheit and Celsius. A thermometer is calibrated by measuring the difference in height between mercury at the freezing point of water, and mercury at the boiling point of water. The interval between these two points is then divided into equal increments—180, as we have seen, for the Fahrenheit scale, and 100 for the Celsius scale.

ELECTRIC THERMOMETERS. Faster temperature measures can be obtained by thermometers using electricity. All matter displays a certain resistance to electrical current, a resistance that changes with temperature. Therefore, a resistance thermometer uses a fine wire wrapped around an insulator, and when a change in temperature occurs, the resistance in the wire changes as well. This makes possible much quick-

KEY TERMS CONTINUED

MOLECULAR TRANSLATIONAL ENERGY: The kinetic energy in a system produced by the movement of molecules in relation to one another.

SYSTEM: In physics, the term "system" usually refers to any set of physical interactions, or any material body, isolated from the rest of the universe. Anything outside of the system, including all factors and forces irrelevant to a discussion of that system, is known as the environment.

TEMPERATURE: A measure of the average kinetic energy—or molecular translational energy in a system. Differences in temperature determine the direction of internal energy flow between two systems when heat is being transferred.

THERMAL ENERGY: Heat energy, a form of kinetic energy produced by the movement of atomic or molecular particles. The greater the movement of these particles, the greater the thermal energy.

THERMAL EQUILIBRIUM: The state that exists when two systems have the same temperature. As a result, there is no exchange of heat between them.

THERMODYNAMICS: The study of the relationships between heat, work, and energy.

THERMOMETRIC MEDIUM: A substance whose properties change with temperature. A mercury or alcohol thermometer measures such changes.

THERMOMETER: A device that gauges temperature by measuring a temperature-dependent property, such as the expansion of a liquid in a sealed tube, or resistance to electric current.

TRIPLE POINT: The temperature and pressure at which a substance is at once a solid, liquid, and vapor.

VACUUM: Space entirely devoid of matter, including air.

er temperature readings than those offered by a thermometer containing a traditional thermometric medium.

Resistance thermometers are highly reliable, but expensive, and are used primarily for very precise measurements. More practical for everyday use is a thermistor, which also uses the principle of electric resistance, but is much simpler and less expensive. Thermistors are used for providing measurements of the internal temperature of food, for instance, and for measuring human body temperature.

Another electric temperature-measurement device is a thermocouple. When wires of two different materials are connected, this creates a small level of voltage that varies as a function of temperature. A typical thermocouple uses two junctions: a reference junction, kept at some constant temperature, and a measurement junction. The measurement junction is applied to the item whose temperature is to be measured, and any temperature difference between it and the reference junction registers as a voltage change, which is measured with a meter connected to the system.

OTHER TYPES OF THERMOMETER. A pyrometer also uses electromagnetic properties, but of a very different kind. Rather than responding to changes in current or voltage, the pyrometer is a gauge that responds to visible and infrared radiation. Temperature and color are closely related: thus, it is no accident that greens, blues, and purples, at one end of the visible light spectrum, are associated with coolness, while reds, oranges, and yellows at the other end are associated with heat. As with the thermocouple, a pyrometer has both a reference element and a measurement element, which compares light readings between the reference filament and the object whose temperature is being measured.

Still other thermometers, such as those in an oven that tell the user its internal temperature, are based on the expansion of metals with heat. In fact, there are a wide variety of thermometers, each suited to a specific purpose. A pyrometer, for instance, is good for measuring the temperature of an object that the thermometer itself does not touch.

WHERE TO LEARN MORE

About Temperature (Web site). <http://www.unidata.ucar.edu/staff/blynds/tmp.html> (April 18, 2001).

About Temperature Sensors (Web site). <http://www.temperatures.com> (April 18, 2001).

Gardner, Robert. *Science Projects About Methods of Measuring.* Berkeley Heights, N.J.: Enslow Publishers, 2000.

Maestro, Betsy and Giulio Maestro. *Temperature and You.* New York: Macmillan/McGraw-Hill School Publishing, 1990.

Megaconverter (Web site). <http://www.megaconverter.com> (April 18, 2001).

NPL: National Physics Laboratory: Thermal Stuff: Beginners' Guides (Web site). <http://www.npl.co.uk/npl/cbtm/thermal/stuff/guides.html> (April 18, 2001).

Royston, Angela. *Hot and Cold.* Chicago: Heinemann Library, 2001.

Santrey, Laurence. *Heat.* Illustrated by Lloyd Birmingham. Mahwah, N.J.: Troll Associates, 1985.

Suplee, Curt. *Everyday Science Explained.* Washington, D.C.: National Geographic Society, 1996.

Walpole, Brenda. *Temperature.* Illustrated by Chris Fairclough and Dennis Tinkler. Milwaukee, WI: Gareth Stevens Publishing, 1995.

THERMAL EXPANSION

CONCEPT

Most materials are subject to thermal expansion: a tendency to expand when heated, and to contract when cooled. For this reason, bridges are built with metal expansion joints, so that they can expand and contract without causing faults in the overall structure of the bridge. Other machines and structures likewise have built-in protection against the hazards of thermal expansion. But thermal expansion can also be advantageous, making possible the workings of thermometers and thermostats.

HOW IT WORKS

Molecular Translational Energy

In scientific terms, heat is internal energy that flows from a system of relatively high temperature to one at a relatively low temperature. The internal energy itself, identified as thermal energy, is what people commonly mean when they say "heat." A form of kinetic energy due to the movement of molecules, thermal energy is sometimes called molecular translational energy.

Temperature is defined as a measure of the average molecular translational energy in a system, and the greater the temperature change for most materials, as we shall see, the greater the amount of thermal expansion. Thus, all these aspects of "heat"—heat itself (in the scientific sense), as well as thermal energy, temperature, and thermal expansion—are ultimately affected by the motion of molecules in relation to one another.

MOLECULAR MOTION AND NEWTONIAN PHYSICS. In general, the kinetic energy created by molecular motion can be understood within the framework of classical physics—that is, the paradigm associated with Sir Isaac Newton (1642-1727) and his laws of motion. Newton was the first to understand the physical force known as gravity, and he explained the behavior of objects within the context of gravitational force. Among the concepts essential to an understanding of Newtonian physics are the mass of an object, its rate of motion (whether in terms of velocity or acceleration), and the distance between objects. These, in turn, are all components central to an understanding of how molecules in relative motion generate thermal energy.

The greater the momentum of an object—that is, the product of its mass multiplied by its rate of velocity—the greater the impact it has on another object with which it collides. The greater, also, is its kinetic energy, which is equal to one-half its mass multiplied by the square of its velocity. The mass of a molecule, of course, is very small, yet if all the molecules within an object are in relative motion—many of them colliding and, thus, transferring kinetic energy—this is bound to lead to a relatively large amount of thermal energy on the part of the larger object.

MOLECULAR ATTRACTION AND PHASES OF MATTER. Yet, precisely because molecular mass is so small, gravitational force alone cannot explain the attraction between molecules. That attraction instead must be understood in terms of a second type of force—electromagnetism—discovered by Scottish physicist James Clerk Maxwell (1831-1879). The details of electromagnetic force are not

Because steel has a relatively high coefficient of thermal expansion, standard railroad tracks are constructed so that they can safely expand on a hot day without derailing the trains traveling over them. *(Milepost 92 1/2/Corbis. Reproduced by permission.)*

important here; it is necessary only to know that all molecules possess some component of electrical charge. Since like charges repel and opposite charges attract, there is constant electromagnetic interaction between molecules, and this produces differing degrees of attraction.

The greater the relative motion between molecules, generally speaking, the less their attraction toward one another. Indeed, these two aspects of a material—relative attraction and motion at the molecular level—determine whether that material can be classified as a solid, liquid, or gas. When molecules move slowly in relation to one another, they exert a strong attraction, and the material of which they are a part is usually classified as a solid. Molecules of liquid, on the other hand, move at moderate speeds, and therefore exert a moderate attraction. When molecules move at high speeds, they exert little or no attraction, and the material is known as a gas.

Predicting Thermal Expansion

COEFFICIENT OF LINEAR EXPANSION. A coefficient is a number that serves as a measure for some characteristic or property. It may also be a factor against which other values are multiplied to provide a desired result. For any type of material, it is possible to

A man ice fishing in Montana. Because of the unique thermal expansion properties of water, ice forms at the top of a lake rather than the bottom, thus allowing marine life to continue living below its surface during the winter. *(Corbis. Reproduced by permission.)*

calculate the degree to which that material will expand or contract when exposed to changes in temperature. This is known, in general terms, as its coefficient of expansion, though, in fact, there are two varieties of expansion coefficient.

The coefficient of linear expansion is a constant that governs the degree to which the length of a solid will change as a result of an alteration in temperature For any given substance, the coefficient of linear expansion is typically a number expressed in terms of 10^{-5}/°C. In other words, the value of a particular solid's linear expansion coefficient is multiplied by 0.00001 per °C. (The °C in the denominator, shown in the equation below, simply "drops out" when the coefficient of linear expansion is multiplied by the change in temperature.)

For quartz, the coefficient of linear expansion is 0.05. By contrast, iron, with a coefficient of 1.2, is 24 times more likely to expand or contract as a result of changes in temperature. (Steel has the same value as iron.) The coefficient for aluminum is 2.4, twice that of iron or steel. This means that an equal temperature change will produce twice as much change in the length of a bar of aluminum as for a bar of iron. Lead is among the most expansive solid materials, with a coefficient equal to 3.0.

CALCULATING LINEAR EXPANSION. The linear expansion of a given solid can be calculated according to the formula $\delta L = aL_O\Delta T$. The Greek letter delta (d) means "a change in"; hence, the first figure represents change in length, while the last figure in the equation stands for change in temperature. The letter *a* is the coefficient of linear expansion, and L_O is the original length.

Suppose a bar of lead 5 meters long experiences a temperature change of 10°C; what will its change in length be? To answer this, *a* ($3.0 \cdot 10^{-5}$/°C) must be multiplied by L_O (5 m) and δT (10°C). The answer should be 150 & 10^{-5} m, or 1.5 mm. Note that this is simply a change in length related to a change in temperature: if the temperature is raised, the length will increase, and if the temperature is lowered by 10°C, the length will decrease by 1.5 mm.

VOLUME EXPANSION. Obviously, linear equations can only be applied to solids. Liquids and gases, classified together as fluids, conform to the shape of their container; hence, the "length" of any given fluid sample is the same as that of the solid that contains it. Fluids are, however, subject to volume expansion—that is, a change in volume as a result of a change in temperature.

To calculate change in volume, the formula is very much the same as for change in length; only a few particulars are different. In the formula $\delta V = bV_O\delta T$, the last term, again, means change in temperature, while δV means change in volume and V_O is the original volume. The letter *b* refers to the coefficient of volume expansion. The latter is expressed in terms of 10^{-4}/°C, or 0.0001 per °C.

Glass has a very low coefficient of volume expansion, 0.2, and that of Pyrex glass is extremely low—only 0.09. For this reason, items made of Pyrex are ideally suited for cooking. Significantly higher is the coefficient of volume expansion for glycerin, an oily substance associated with soap, which expands proportionally to a factor of 5.1. Even higher is ethyl alcohol, with a volume expansion coefficient of 7.5.

REAL-LIFE APPLICATIONS

Liquids

Most liquids follow a fairly predictable pattern of gradual volume increase, as a response to an increase in temperature, and volume decrease, in response to a decrease in temperature. Indeed, the coefficient of volume expansion for a liquid generally tends to be higher than for a solid, and—with one notable exception discussed below— a liquid will contract when frozen.

The behavior of gasoline pumped on a hot day provides an example of liquid thermal expansion in response to an increase in temperature. When it comes from its underground tank at the gas station, the gasoline is relatively cool, but it will warm when sitting in the tank of an already warm car. If the car's tank is filled and the vehicle left to sit in the sun—in other words, if the car is not driven after the tank is filled—the gasoline might very well expand in volume faster than the fuel tank, overflowing onto the pavement.

ENGINE COOLANT. Another example of thermal expansion on the part of a liquid can be found inside the car's radiator. If the radiator is "topped off" with coolant on a cold day, an increase in temperature could very well cause the coolant to expand until it overflows. In the past, this produced a problem for car owners, because car engines released the excess volume of coolant onto the ground, requiring periodic replacement of the fluid.

Later-model cars, however, have an overflow container to collect fluid released as a result of volume expansion. As the engine cools down again, the container returns the excess fluid to the radiator, thus, "recycling" it. This means that newer cars are much less prone to overheating as older cars. Combined with improvements in radiator fluid mixtures, which act as antifreeze in cold weather and coolant in hot, the "recycling" process has led to a significant decrease in breakdowns related to thermal expansion.

WATER. One good reason not to use pure water in one's radiator is that water has a far higher coefficient of volume expansion than a typical engine coolant. This can be particularly hazardous in cold weather, because frozen water in a radiator could expand enough to crack the engine block.

In general, water—whose volume expansion coefficient in the liquid state is 2.1, and 0.5 in the solid state—exhibits a number of interesting characteristics where thermal expansion is concerned. If water is reduced from its boiling point—212°F (100°C) to 39.2°F (4°C) it will steadily contract, like any other substance responding to a drop in temperature. Normally, however, a substance continues to become denser as it turns from liquid to solid; but this does not occur with water.

At 32.9°F, water reaches it maximum density, meaning that its volume, for a given unit of mass, is at a minimum. Below that temperature, it "should" (if it were like most types of matter) continue to decrease in volume per unit of mass, but, in fact, it steadily begins to expand. Thus, it is less dense, with a greater volume per unit of mass, when it reaches the freezing point. It is for this reason that when pipes freeze in winter, they often burst—explaining why a radiator filled with water could be a serious problem in very cold weather.

In addition, this unusual behavior with regard to thermal expansion and contraction explains why ice floats: solid water is less dense than the liquid water below it. As a result, frozen water stays at the top of a lake in winter; since ice is a poor conductor of heat, energy cannot escape from the water below it in sufficient amounts to freeze the rest of the lake water. Thus, the water below the ice stays liquid, preserving plant and animal life.

Gases

THE GAS LAWS. As discussed, liquids expand by larger factors than solids do. Given the increasing amount of molecular kinetic energy for a liquid as compared to a solid, and for a gas as compared to a liquid, it should not be surprising, then, to learn that gases respond to changes in temperature with a volume change even greater than that of liquids. Of course, where a gas is concerned, "volume" is more difficult to measure, because a gas simply expands to fill its container. In order for the term to have any meaning, pressure and temperature must be specified as well.

A number of the gas laws describe the three parameters for gases: volume, temperature, and pressure. Boyle's law, for example, holds that in conditions of constant temperature, an inverse relationship exists between the volume and pressure of a gas: the greater the pressure, the less the volume, and vice versa. Even more relevant to the subject of thermal expansion is Charles's law.

Charles's law states that when pressure is kept constant, there is a direct relationship between volume and temperature. As a gas heats up, its volume increases, and when it cools down, its volume reduces accordingly. Thus, if an air mattress is filled in an air-conditioned room, and the mattress is then taken to the beach on a hot day, the air inside will expand. Depending on how much its volume increases, the expansion of the hot air could cause the mattress to "pop."

VOLUME GAS THERMOMETERS. Whereas liquids and solids vary significantly with regard to their expansion coefficients, most gases follow more or less the same pattern of expansion in response to increases in temperature. The predictable behavior of gases in these situations led to the development of the constant gas thermometer, a highly reliable instrument against which other thermometers—including those containing mercury (see below)—are often gauged.

In a volume gas thermometer, an empty container is attached to a glass tube containing mercury. As gas is released into the empty container, this causes the column of mercury to move upward. The difference between the former position of the mercury and its position after the introduction of the gas shows the difference between normal atmospheric pressure and the pressure of the gas in the container. It is, then, possible to use the changes in volume on the part of the gas as a measure of temperature. The response of most gases, under conditions of low pressure, to changes in temperature is so uniform that volume gas thermometers are often used to calibrate other types of thermometers.

Solids

Many solids are made up of crystals, regular shapes composed of molecules joined to one another as though on springs. A spring that is pulled back, just before it is released, is an example of potential energy, or the energy that an object possesses by virtue of its position. For a crystalline solid at room temperature, potential energy and spacing between molecules are relatively low. But as temperature increases and the solid expands, the space between molecules increases—as does the potential energy in the solid.

In fact, the responses of solids to changes in temperature tend to be more dramatic, at least when they are seen in daily life, than are the behaviors of liquids or gases under conditions of thermal expansion. Of course, solids actually respond less to changes in temperature than fluids do; but since they are solids, people expect their contours to be immovable. Thus, when the volume of a solid changes as a result of an increase in thermal energy, the outcome is more noteworthy.

JAR LIDS AND POWER LINES. An everyday example of thermal expansion can be seen in the kitchen. Almost everyone has had the experience of trying unsuccessfully to budge a tight metal lid on a glass container, and after running hot water over the lid, finding that it gives way and opens at last. The reason for this is that the high-temperature water causes the metal lid to expand. On the other hand, glass—as noted earlier—has a low coefficient of expansion. Otherwise, it would expand with the lid, which would defeat the purpose of running hot water over it. If glass jars had a high coefficient of expansion, they would deform when exposed to relatively low levels of heat.

Another example of thermal expansion in a solid is the sagging of electrical power lines on a hot day. This happens because heat causes them to expand, and, thus, there is a greater length of power line extending from pole to pole than under lower temperature conditions. It is highly unlikely, of course, that the heat of summer could be so great as to pose a danger of power lines breaking; on the other hand, heat can create a serious threat with regard to larger structures.

EXPANSION JOINTS. Most large bridges include expansion joints, which look rather like two metal combs facing one another, their teeth interlocking. When heat causes the bridge to expand during the sunlight hours of a hot day, the two sides of the expansion joint move toward one another; then, as the bridge cools down after dark, they begin gradually to retract. Thus the bridge has a built-in safety zone; otherwise, it would have no room for expansion or contraction in response to temperature changes. As for the use of the comb shape, this staggers the gap between the two sides of the expansion joint, thus minimizing the bump motorists experience as they drive over it.

Expansion joints of a different design can also be found in highways, and on "highways" of rail. Thermal expansion is a particularly serious problem where railroad tracks are concerned, since the tracks on which the trains run are made of steel. Steel, as noted earlier, expands by a factor of 12 parts in 1 million for every Celsius degree change in temperature, and while this may not seem like much, it can create a serious problem under conditions of high temperature.

Most tracks are built from pieces of steel supported by wooden ties, and laid with a gap between the ends. This gap provides a buffer for thermal expansion, but there is another matter to consider: the tracks are bolted to the wooden ties, and if the steel expands too much, it could pull out these bolts. Hence, instead of being placed in a hole the same size as the bolt, the bolts are fitted in slots, so that there is room for the track to slide in place slowly when the temperature rises.

Such an arrangement works agreeably for trains that run at ordinary speeds: their wheels merely make a noise as they pass over the gaps, which are rarely wider than 0.5 in (0.013 m). A high-speed train, however, cannot travel over irregular track; therefore, tracks for high-speed trains are laid under conditions of relatively high tension. Hydraulic equipment is used to pull sections of the track taut; then, once the track is secured in place along the cross ties, the tension is distributed down the length of the track.

Thermometers and Thermostats

MERCURY IN THERMOMETERS. A thermometer gauges temperature by measuring a temperature-dependent property. A thermostat, by contrast, is a device for adjusting the temperature of a heating or cooling system. Both use the principle of thermal expansion in their operation. As noted in the example of the metal lid and glass jar above, glass expands little with changes in temperature; therefore, it makes an ideal container for the mercury in a thermometer. As for mercury, it is an ideal thermometric medium—that is, a material used to gauge temperature—for several reasons. Among these is a high boiling point, and a highly predictable, uniform response to changes in temperature.

In a typical mercury thermometer, mercury is placed in a long, narrow sealed tube called a capillary. Because it expands at a much faster rate than the glass capillary, mercury rises and falls with the temperature. A thermometer is calibrated by measuring the difference in height between mercury at the freezing point of water, and mercury at the boiling point of water. The interval

KEY TERMS

COEFFICIENT: A number that serves as a measure for some characteristic or property. A coefficient may also be a factor against which other values are multiplied to provide a desired result.

COEFFICIENT OF LINEAR EXPANSION: A figure, constant for any particular type of solid, used in calculating the amount by which the length of that solid will change as a result of temperature change. For any given substance, the coefficient of linear expansion is typically a number expressed in terms of 10^{-5}/°C.

COEFFICIENT OF VOLUME EXPANSION: A figure, constant for any particular type of material, used in calculating the amount by which the volume of that material will change as a result of temperature change. For any given substance, the coefficient of volume expansion is typically a number expressed in terms of 10^{-4}/°C.

HEAT: Internal thermal energy that flows from one body of matter to another.

KINETIC ENERGY: The energy that an object possesses by virtue of its motion.

MOLECULAR TRANSLATIONAL ENERGY: The kinetic energy in a system produced by the movement of molecules in relation to one another.

POTENTIAL ENERGY: The energy that an object possesses by virtue of its position.

SYSTEM: In physics, the term "system" usually refers to any set of physical interactions, or any material body, isolated from the rest of the universe. Anything outside of the system, including all factors and forces irrelevant to a discussion of that system, is known as the environment.

TEMPERATURE: A measure of the average kinetic energy—or molecular translational energy in a system. Differences in temperature determine the direction of internal energy flow between two systems when heat is being transferred.

THERMAL ENERGY: Heat energy, a form of kinetic energy produced by the movement of atomic or molecular particles. The greater the movement of these particles, the greater the thermal energy.

THERMAL EXPANSION: A property in all types of matter that display a tendency to expand when heated, and to contract when cooled.

between these two points is then divided into equal increments in accordance with one of the well-known temperature scales.

THE BIMETALLIC STRIP IN THERMOSTATS. In a thermostat, the central component is a bimetallic strip, consisting of thin strips of two different metals placed back to back. One of these metals is of a kind that possesses a high coefficient of linear expansion, while the other metal has a low coefficient. A temperature increase will cause the side with a higher coefficient to expand more than the side that is less responsive to temperature changes. As a result, the bimetallic strip will bend to one side.

When the strip bends far enough, it will close an electrical circuit, and, thus, direct the air conditioner to go into action. By adjusting the thermostat, one varies the distance that the bimetallic strip must be bent in order to close the circuit. Once the air in the room reaches the desired temperature, the high-coefficient metal will begin to contract, and the bimetallic strip will straighten. This will cause an opening of the electrical circuit, disengaging the air conditioner.

In cold weather, when the temperature-control system is geared toward heating rather than cooling, the bimetallic strip acts in much the same way—only this time, the high-coefficient metal contracts with cold, engaging the heater. Another type of thermostat uses the expansion of a vapor rather than a solid. In this case, heating of the vapor causes it to expand, pushing on a set of brass bellows and closing the circuit, thus, engaging the air conditioner.

WHERE TO LEARN MORE

Beiser, Arthur. *Physics,* 5th ed. Reading, MA: Addison-Wesley, 1991.

"*Comparison of Materials: Coefficient of Thermal Expansion*" (Web site). <http://www.handyharmancanada.com/TheBrazingBook/comparis.html> (April 21, 2001).

Encyclopedia of Thermodynamics (Web site). <http://therion.minpet.unibas.ch/minpet/groups/thermodict/> (April 12, 2001).

Fleisher, Paul. *Matter and Energy: Principles of Matter and Thermodynamics.* Minneapolis, MN: Lerner Publications, 2002.

NPL: National Physics Laboratory: Thermal Stuff: Beginners' Guides (Web site). <http://www.npl.co.uk/npl/cbtm/thermal/stuff/guides.html> (April 18, 2001).

Royston, Angela. *Hot and Cold.* Chicago: Heinemann Library, 2001.

Suplee, Curt. *Everyday Science Explained.* Washington, D.C.: National Geographic Society, 1996.

"*Thermal Expansion Measurement*" (Web site). <http://www.measurementsgroup.com/guide/tn/tn513/513intro.html> (April 21, 2001).

"*Thermal Expansion of Solids and Liquids*" (Web site). <http://www.physics.mun.ca/~gquirion/P2053/html19b/> (April 21, 2001).

Walpole, Brenda. *Temperature.* Illustrated by Chris Fairclough and Dennis Tinkler. Milwaukee, WI: Gareth Stevens Publishing, 1995.

SCIENCE OF EVERYDAY THINGS
REAL-LIFE PHYSICS

WAVE MOTION AND OSCILLATION

WAVE MOTION

OSCILLATION

FREQUENCY

RESONANCE

INTERFERENCE

DIFFRACTION

DOPPLER EFFECT

WAVE MOTION

CONCEPT

Wave motion is activity that carries energy from one place to another without actually moving any matter. Studies of wave motion are most commonly associated with sound or radio transmissions, and, indeed, these are among the most common forms of wave activity experienced in daily life. Then, of course, there are waves on the ocean or the waves produced by an object falling into a pool of still water—two very visual examples of a phenomenon that takes place everywhere in the world around us.

HOW IT WORKS

Related Forms of Motion

In wave motion, energy—the ability to perform work, or to exert force over distance—is transmitted from one place to another without actually moving any matter along the wave. In some types of waves, such as those on the ocean, it might seem as though matter itself has been displaced; that is, it appears that the water has actually moved from its original position. In fact, this is not the case: molecules of water in an ocean wave move up and down, but they do not actually travel with the wave itself. Only the energy is moved.

A wave is an example of a larger class of regular, repeated, and/or back-and-forth types of motion. As with wave motion, these varieties of movement may or may not involve matter, but, in any case, the key component is not matter, but energy. Broadest among these is periodic motion, or motion that is repeated at regular intervals called periods. A period might be the amount of time that it takes an object orbiting another (as, for instance, a satellite going around Earth) to complete one cycle of orbit. With wave motion, a period is the amount of time required to complete one full cycle of the wave, from trough to crest and back to trough.

HARMONIC MOTION. Harmonic motion is the repeated movement of a particle about a position of equilibrium, or balance. In harmonic motion—or, more specifically, simple harmonic motion—the object moves back and forth under the influence of a force directed toward the position of equilibrium, or the place where the object stops if it ceases to be in motion. A familiar example of harmonic motion, to anyone who has seen an old movie with a clichéd depiction of a hypnotist, is the back-and-forth movement of the hypnotist's watch, as he tries to control the mind of his patient.

One variety of harmonic motion is vibration, which wave motion resembles in some respects. Both wave motion and vibration are periodic, involving the regular repetition of a certain form of movement. In both, there is a continual conversion and reconversion between potential energy (the energy of an object due to its position, as for instance with a sled at the top of a hill) and kinetic energy (the energy of an object due to its motion, as with the sled when sliding down the hill.) The principal difference between vibration and wave motion is that, in the first instance, the energy remains in place, whereas waves actually transport energy from one place to another.

OSCILLATION. Oscillation is a type of harmonic motion, typically periodic, in one or more dimensions. Suppose a spring is fixed in

Heinrich Hertz. *(Hulton-Deutsch Collection/Corbis. Reproduced by permission.)*

place to a ceiling, such that it hangs downward. At this point, the spring is in a position of equilibrium. Now, consider what happens if the spring is grasped at a certain point and lifted, then let go. It will, of course, fall downward with the force of gravity until it comes to a stop—but it will not stop at the earlier position of equilibrium. Instead, it will continue downward to a point of maximum tension, where it possesses maximum potential energy as well. Then, it will spring upward again, and as it moves, its kinetic energy increases, while potential energy decreases. At the high point of this period of oscillation, the spring will not be as high as it was before it was originally released, but it will be higher than the position of equilibrium.

Once it falls, the spring will again go lower than the position of equilibrium, but not as low as before—and so on. This is an example of oscillation. Now, imagine what happens if another spring is placed beside the first one, and they are connected by a rubber band. If just the first spring is disturbed, as before, the second spring will still move, because the energy created by the movement of the first spring will be transmitted to the second one via the rubber band. The same will happen if a row of springs, all side-by-side, are attached by multiple rubber bands, and the first spring is once again disturbed: the energy will pass through the rubber bands, from spring to spring, causing the entire row to oscillate. This is similar to what happens in the motion of a wave.

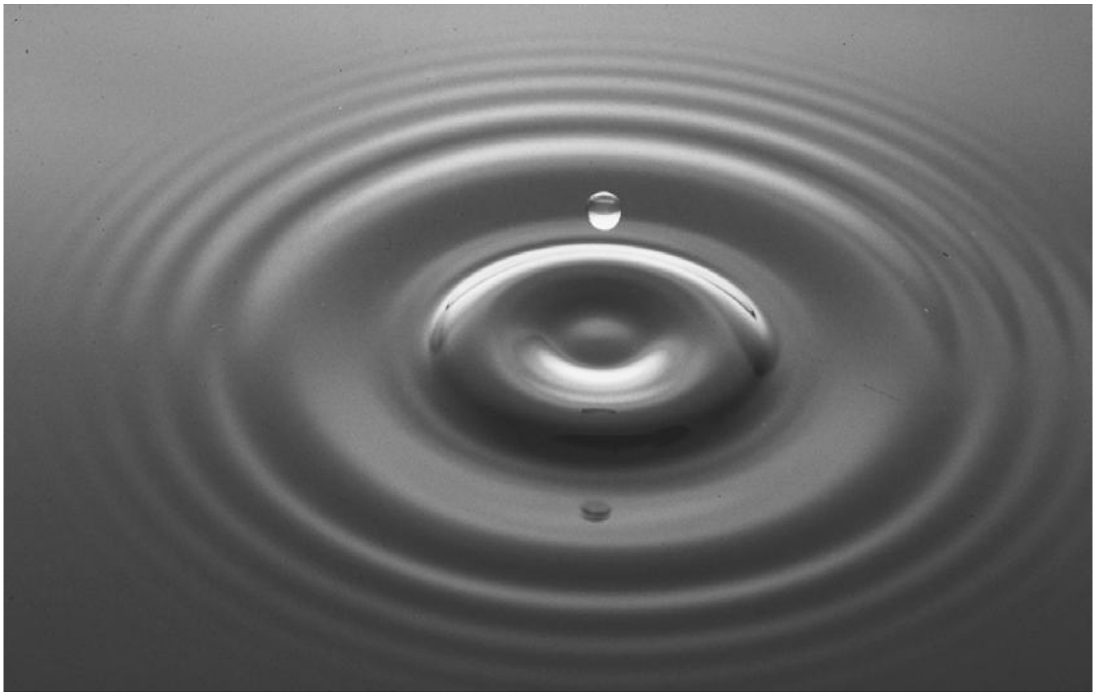
Types and Properties of Waves

There are some types of waves that do not follow regular, repeated patterns; these are discussed below, in the illustration concerning a string, in which a pulse is created and reflected. Of principal concern here, however, is the periodic wave, a series of wave motions, following one after the other in regular succession. Examples of periodic waves include waves on the ocean, sound waves, and electromagnetic waves. The last of these include visible light and radio, among others.

Electromagnetic waves involve only energy; on the other hand, a mechanical wave involves matter as well. Ocean waves are mechanical waves; so, too, are sound waves, as well as the waves produced by pulling a string. It is important to note, again, that the matter itself is not moved from place to place, though it may move in place without leaving its position. For example, water molecules in the crest of an ocean wave rotate in the same direction as the wave, while those in the trough of the wave rotate in a direction opposite to that of the wave, yet there is no net motion of the water: only energy is transmitted along the wave.

FIVE PROPERTIES OF WAVES. There are three notable interrelated characteristics of periodic waves. One of these is wave speed, symbolized by v and typically calculated in meters per second. Another is wavelength, represented as λ (the Greek letter lambda), which is the distance between a crest and the adjacent crest, or a trough and the adjacent trough. The third is frequency, abbreviated as f, which is the number of waves passing through a given point during the interval of 1 second.

Frequency is measured in terms of cycles per second, or Hertz (Hz), named in honor of nineteenth-century German physicist Heinrich Rudolf Hertz (1857-1894). If a wave has a frequency of 100 Hz, this means that 100 waves are passing through a given point during the interval of 1 second. Higher frequencies are expressed in terms of kilohertz (kHz; 10^3 or 1,000 cycles per



Transverse waves produced by a water droplet penetrating the surface of a body of liquid. *(Photograph by Martin Dohrn/Science Photo Library, National Audubon Society Collection/Photo Researchers, Inc. Reproduced with permission.)*

second) or megahertz (MHz; 10^6 or 1 million cycles per second.)

Frequency is clearly related to wave speed, and there is also a relationship—though it is not so immediately grasped—between wavelength and speed. Over the interval of 1 second, a given number of waves pass a certain point (frequency), and each wave occupies a certain distance (wavelength). Multiplied by one another, these two properties equal the speed of the wave. This can be stated as a formula: $v = f\lambda$.

Earlier, the term "period" was defined in terms of wave motion as the amount of time required to complete one full cycle of the wave. Period, symbolized by T, can be expressed in terms of frequency, and, thus, can also be related to the other two properties identified above. It is the inverse of frequency, meaning that $T = 1/f$. Furthermore, period is equal to the ratio of wavelength to wave speed; in other words, $T = \lambda/v$.

A fifth property of waves—one not mathematically related to wavelength, wave speed, frequency, or period, is amplitude. Amplitude can be defined as the maximum displacement of oscillating particles from their normal position. For an ocean wave, amplitude is the distance from either the crest or the trough to the level that the ocean would maintain if it were perfectly still.

WAVE SHAPES. When most people think of waves, naturally, one of the first images that comes to mind is that of waves on the ocean. These are an example of a transverse wave, or one in which the vibration or motion is perpendicular to the direction the wave is moving. (Actually, ocean waves are simply perceived as transverse waves; in fact, as discussed below, their behavior is rather more complicated.) In a longitudinal wave, on the other hand, the movement of vibration is in the same direction as the wave itself.

Transverse waves are easier to visualize, particularly with regard to the aspects of wave motion—for example, frequency and amplitude—discussed above. Yet, longitudinal waves can be understood in terms of a common example. Sound waves, for instance, are longitudinal: thus, when a stereo is turned up to a high volume, the speakers vibrate in the same direction as the sound itself.

A longitudinal wave may be understood as a series of fluctuations in density. If one were to take a coiled spring (such as the toy known as the "Slinky") and release one end while holding the other, the motion of the springs would produce longitudinal waves. As these waves pass through the spring, they cause some portions of it to be compressed and others extended. The distance between each point of compression is the wavelength.

Now, to return to the qualified statement made above: that ocean waves are an example of transverse waves. We perceive them as transverse waves, but, in fact, they are also longitudinal. In fact, all types of waves on the surface of a liquid are a combination of longitudinal and transverse, and are known as surface waves. Thus, if one drops a stone into a body of still water, waves radiate outward (longitudinal), but these waves also have a component that is perpendicular to the surface of the water, meaning that they are also transverse.

REAL-LIFE APPLICATIONS

Pulses on a String

There is another variety of wave, though it is defined in terms of behavior rather than the direction of disturbance. (In terms of direction, it is simply a variety of transverse wave.) This is a standing wave, produced by causing vibrations on a string or other piece of material whose ends are fixed in place. Standing waves are really a series of pulses that travel down the string and are reflected back to the point of the original disturbance.

Suppose you hold a string in one hand, with the other end attached to a wall. If you give the string a shake, this causes a pulse—an isolated, non-periodic disturbance—to move down it. A pulse is a single wave, and the behavior of this lone wave helps us to understand what happens within the larger framework of wave motion. As with wave motion in general, the movement of the pulse involves both kinetic and potential energy. The tension of the string itself creates potential energy; then, as the movement of the pulse causes the string to oscillate upward and downward, this generates a certain amount of kinetic energy.

TENSION AND REFLECTION. The speed of the pulse is a function of the string and its properties, not of the way that the pulse was originally delivered. The tighter the string, and the less its mass per unit of length, the faster the pulse travels down it. The greater the mass per unit of length, however, the greater the inertia resisting the movement of the pulse. Furthermore, the more loosely you hold the string, the less it will respond to the movement of the pulse.

In accordance with the third law of motion, there should be an equal and opposite reaction once the pulse comes into contact with the wall. Assuming that you are holding the string tightly, this reaction will be manifested in the form of an inverted wave, or one that is upside-down in relation to the original pulse. In this case, the tension on the end attached to the support is equal and opposite to the tension exerted by your hand. As a result, the pulse comes back in the same shape as before, but inverted.

If, on the other hand, you hold the other end of the string loosely; instead, once it reaches the wall, its kinetic energy will be converted into potential energy, which will cause the end of the string closest to the wall to move downward. This will result in sending back a pulse that is reversed in horizontal direction, but the same in vertical direction.

In both cases, the energy in the string is reflected backward to its source—that is, to the place from which the pulse was originally produced by the action of your hand. If, however, you hold the string so that its level of tension is exactly between perfect rigidity and perfect looseness, then the pulse will not be reflected. In other words, there will be no reflected wave.

TRANSMISSION AND REFLECTION. If two strings are joined end-to-end, and a pulse is produced at one end, the pulse would, of course, be transmitted to the second string. If, however, the second string has a greater mass per unit of length than the first one, the result would be two pulses: a transmitted pulse moving in the "right" direction, and a reflected, inverted pulse, moving toward the original source of energy. If, on the other hand, the first string has a greater mass per unit of length than the second one, the reflected pulse would be erect (right side up), not inverted.

For simplicity's sake, this illustration has been presented in terms of a string attached to a wall, but, in fact, transmission and reflection occur in a number of varieties of wave motion—

not just those involving pulses or standing waves. A striking example occurs when light hits an ordinary window. The majority of the light, of course, is transmitted through the window pane, but a portion is reflected. Thus, as one looks through the window, one also sees one's reflection.

Similarly, sound waves are reflected depending on the medium with which they are in contact. A canyon wall, for instance, will reflect a great deal of sound, and, thus, it is easy to produce an echo in such a situation. On the other hand, there are many instances in which the desire is to "absorb" sound by transmitting it to some other form of material. Thus, for example, the lobby of an upscale hotel will include a number of plants, as well as tapestries and various wall hangings. In addition to adding beauty, these provide a medium into which the sound of voices and other noises can be transmitted and, thus, absorbed.

Sound Waves

PRODUCTION. The experience of sound involves production, or the generation of sound waves; transmission, or the movement of those waves from their source; and reception, the principal example of which is hearing. Sound itself is discussed in detail elsewhere. Of primary concern here is the transmission, and to a lesser extent, the production of sound waves.

In terms of production, sound waves are, as noted, longitudinal waves: changes in pressure, or alternations between condensation and rarefaction. Vibration is integral to the generation of sound. When the diaphragm of a loudspeaker pushes outward, it forces nearby air molecules closer together, creating a high-pressure region all around the loudspeaker. The loudspeaker's diaphragm is pushed backward in response, thus freeing up a volume of space for the air molecules. These, then, rush toward the diaphragm, creating a low-pressure region behind the high-pressure one. As a result, the loudspeaker sends out alternating waves of high pressure (condensation) and low pressure (rarefaction).

FREQUENCY AND WAVELENGTH. As sound waves pass through a medium such as air, they create fluctuations between condensation and rarefaction. These result in pressure changes that cause the listener's eardrum to vibrate with the same frequency as the sound wave, a vibration that the ear's inner mechanisms translate and pass on to the brain. The range of audibility for the human ear is from 20 Hz to 20 kHz. The lowest note of the eighty-eight keys on a piano is 27 Hz and the highest 4.186 kHz. This places the middle and upper register of the piano well within the optimal range for audibility, which is between 3 and 4 kHz.

Sound travels at a speed of about 1,088 ft (331 m) per second through air at sea level, and the range of sound audible to human ears includes wavelengths as large as 11 ft (3.3 m) and as small as 1.3 in (3.3 cm). Unlike light waves, which are very small, the wavelengths of audible sound are comparable to the sizes of ordinary objects. This creates an interesting contrast between the behaviors of sound and light when confronted with an obstacle to their transmission.

It is fairly easy to block out light by simply holding up a hand in front of one's eyes. When this happens, the Sun casts a shadow on the other side of one's hand. The same action does not work with one's ears and the source of a sound, however, because the wavelengths of sound are large enough to go right past a relatively small object such as a hand. However, if one were to put up a tall, wide cement wall between oneself and the source of a sound—as is often done in areas where an interstate highway passes right by a residential community—the object would be sufficiently large to block out much of the sound.

Radio Waves

Radio waves, like visible light waves, are part of the electromagnetic spectrum. They are characterized by relatively long wavelengths and low frequencies—low, that is, in contrast to the much higher frequencies of both visible and invisible light waves. The frequency range of radio is between 10 KHz and about 2,000 MHz—in other words, from 10,000 Hz to as much as 2 billion Hz—an impressively wide range.

AM radio broadcasts are found between 0.6 and 1.6 MHz, and FM broadcasts between 88 and 108 MHz. Thus, FM is at a much, much higher frequency than AM, with the lowest frequency on the FM dial 55 times as great as the highest on the AM dial. There are other ranges of frequency assigned by the FCC (Federal Communications Commission) to other varieties of radio trans-

KEY TERMS

AMPLITUDE: The maximum displacement of particles in oscillation from their normal position. For an ocean wave, amplitude is the distance from either the crest or the trough to the level that the ocean would maintain if it were perfectly still.

ENERGY: The ability to perform work, which is the exertion of force over a given distance. Work is the product of force and distance, where force and distance are exerted in the same direction.

FREQUENCY: The number of waves passing through a given point during the interval of one second. The higher the frequency, the shorter the wavelength. Frequency can also be mathematically related to wave speed and period.

HARMONIC MOTION: The repeated movement of a particle about a position of equilibrium, or balance.

HERTZ: A unit for measuring frequency, equal to one cycle per second. If a sound wave has a frequency of 20,000 Hz, this means that 20,000 waves are passing through a given point during the interval of one second. Higher frequencies are expressed in terms of kilohertz (kHz; 10^3 or 1,000 cycles per second) or megahertz (MHz; 10^6 or 1 million cycles per second).

KINETIC ENERGY: The energy that an object possesses due to its motion, as with a sled when sliding down a hill. This is contrasted with potential energy.

LONGITUDINAL WAVE: A wave in which the movement of vibration is in the same direction as the wave itself. This is contrasted to a transverse wave.

MATTER: Physical substance that has mass; occupies space; is composed of atoms; and is ultimately convertible to energy.

MECHANICAL WAVE: A type of wave that involves matter. Ocean waves are mechanical waves; so, too, are the waves produced by pulling a string. The matter itself may move in place, but, as with all

mission: for instance, citizens' band (CB) radios are in a region between AM and FM, ranging from 26.985 MHz to 27.405 MHz.

Frequency does not indicate power. The power of a radio station is a function of the wattage available to its transmitter: hence, radio stations often promote themselves with announcements such as "operating with 100,000 watts of power...." Thus, an AM station, though it has a much lower frequency than an FM station, may possess more power, depending on the wattage of the transmitter. Indeed, as we shall see, it is precisely because of its high frequency that an FM station lacks the broadcast range of an AM station.

AMPLITUDE AND FREQUENCY MODULATIONS. What is the difference between AM and FM? Or to put it another way, why is it that an AM station may be heard halfway across the country, yet its sound on a car radio fades out when the car goes under an overpass? The difference relates to how the various radio signals are modulated.

A radio signal is simply a carrier: it may carry Morse code, or it may carry complex sounds, but in order to transmit voices and music, its signal must be modulated. This can be done, for instance, by varying the instantaneous amplitude of the radio wave, which is a function of the radio station's power. These variations in amplitude are called amplitude modulation, or

KEY TERMS CONTINUED

types of wave motion, there is no net movement of matter—only of energy.

OSCILLATION: A type of harmonic motion, typically periodic, in one or more dimensions.

PERIOD: For wave motion, a period is the amount of time required to complete one full cycle of the wave, from trough to crest and back to trough. Period can be mathematically related to frequency, wavelength, and wave speed.

PERIODIC MOTION: Motion that is repeated at regular intervals. These intervals are known as periods.

PERIODIC WAVE: A wave in which a uniform series of crests and troughs follow one after the other in regular succession. By contrast, the wave produced by applying a pulse to a stretched string does not follow regular, repeated patterns.

POTENTIAL ENERGY: The energy that an object possesses due to its position, as for instance with a sled at the top of a hill. This is contrasted with kinetic energy.

PULSE: An isolated, non-periodic disturbance that takes place in wave motion of a type other than that of a periodic wave.

STANDING WAVE: A type of transverse wave produced by causing vibrations on a string or other piece of material whose ends are fixed in place.

SURFACE WAVE: A wave that exhibits the behavior of both a transverse wave and a longitudinal wave.

TRANSVERSE WAVE: A wave in which the vibration or motion is perpendicular to the direction in which the wave is moving. This is contrasted to a longitudinal wave.

WAVELENGTH: The distance between a crest and the adjacent crest, or the trough and an adjacent trough, of a wave. Wavelength, abbreviated λ (the Greek letter lambda) is mathematically related to wave speed, period, and frequency.

WAVE MOTION: Activity that carries energy from one place to another without actually moving any matter.

AM, and this was the first type of commercial radio to appear. Developed in the period before World War I, AM made its debut as a popular phenomenon shortly after the war.

Ironically, FM (frequency modulation) was developed not long after AM, but it did not become commercially viable until well after World War II. As its name suggests, frequency modulation involves variation in the signal's frequency. The amplitude stays the same, and this—combined with the high frequency—produces a nice, even sound for FM radio.

But the high frequency also means that FM signals do not travel as far. If a person is listening to an FM station while moving away from the station's signal, eventually the station will be below the horizon relative to the car, and the car radio will no longer be able to receive the signal. In contrast to the direct, or line-of-sight, transmissions of FM stations, AM signals (with their longer wavelengths) are reflected off of layers in Earth's ionosphere. As a result, a nighttime signal from a "clear channel station" may be heard across much of the continental United States.

WHERE TO LEARN MORE

Ardley, Neil. *Sound Waves to Music.* New York: Gloucester Press, 1990.

Berger, Melvin and Gilda Berger. *What Makes an Ocean Wave?: Questions and Answers About Oceans and Ocean Life.* New York: Scholastic, 2001.

Catherall, Ed. *Exploring Sound.* Austin, TX: Steck-Vaughn Library, 1990.

Glover, David. *Sound and Light.* New York: Kingfisher Books, 1993.

"*Longitudinal and Transverse Wave Motion*" (Web site). <http://www.kettering.edu/~drussell/Demos/waves/wavemotion.html> (April 22, 2001).

"*Multimedia Activities: Wave Motion.*" *ExploreScience.com* (Web site). <http://www.explorescience.com/activities/activity_list.cfm?catego ryID=3> (April 22, 2001).

Ruchlis, Hyman. *Bathtub Physics.* Edited by Donald Barr; illustrated by Ray Skibinski. New York: Harcourt, Brace, and World, 1967.

"*Wave Motion*" (Web site). <http://www.media.uwe.ac.uk/masoud/projects/water/wave.html> (April 22, 2001).

"*Wave Motion and Sound.*" *The Physics Web* (Web site). <http://www.hcrhs.hunterdon.k12.nj.us/disk2/Physics/wave.html> (April 22, 2001).

"*Wave Motion Menu.*"*Carson City-Crystal High School Physics and Chemistry Departments* (Web site). <http://members.aol.com/cepeirce/b2.html> (April 22, 2001).

OSCILLATION

CONCEPT

When a particle experiences repeated movement about a position of stable equilibrium, or balance, it is said to be in harmonic motion, and if this motion is repeated at regular intervals, it is called periodic motion. Oscillation is a type of harmonic motion, typically periodic, in one or more dimensions. Among the examples of oscillation in the physical world are the motion of a spring, a pendulum, or even the steady back-and-forth movement of a child on a swing.

HOW IT WORKS

Stable and Unstable Equilibrium

When a state of equilibrium exists, the vector sum of the forces on an object is equal to zero. There are three varieties of equilibrium: stable, unstable, and neutral. Neutral equilibrium, discussed in the essay on Statics and Equilibrium elsewhere in this book, does not play a significant role in oscillation; on the other hand, stable and unstable equilibrium do.

In the example of a playground swing, when the swing is simply hanging downward—either empty or occupied—it is in a position of stable equilibrium. The vector sums are balanced, because the swing hangs downward with a force (its weight) equal to the force of the bars on the swing set that hold it up. If it were disturbed from this position, as, for instance, by someone pushing the swing, it would tend to return to its original position.

If, on the other hand, the swing were raised to a certain height—if, say, a child were swinging and an adult caught the child at the point of maximum displacement—this would be an example of unstable equilibrium. The swing is in equilibrium because the forces on it are balanced: it is being held upward with a force equal to its weight. Yet, this equilibrium is unstable, because a disturbance (for instance, if the adult lets go of the swing) will cause it to move. Since the swing tends to oscillate, it will move back and forth across the position of stable equilibrium before finally coming to a rest in the stable position.

Properties of Oscillation

There are two basic models of oscillation to consider, and these can be related to the motion of two well-known everyday objects: a spring and a swing. As noted below, objects not commonly considered "springs," such as rubber bands, display spring-like behavior; likewise one could substitute "pendulum" for swing. In any case, it is easy enough to envision the motion of these two varieties of oscillation: a spring generally oscillates along a straight line, whereas a swing describes an arc.

Either case involves properties common to all objects experiencing oscillation. There is always a position of stable equilibrium, and there is always a cycle of oscillation. In a single cycle, the oscillating particle moves from a certain point in a certain direction, then reverses direction and returns to the original point. The amount of time it takes to complete one cycle is called a period, and the number of cycles that take place during one second is the frequency of the oscillation. Frequency is measured in Hertz (Hz), with 1 Hz—the term is both singular and plural—equal to one cycle per second.

The bounce provided by a trampoline is due to elastic potential energy. *(Photograph by Kevin Fleming/Corbis. Reproduced by permission.)*

It is easiest to think of a cycle as the movement from a position of stable equilibrium to one of maximum displacement, or the furthest possible point from stable equilibrium. Because stable equilibrium is directly in the middle of a cycle, there are two points of maximum displacement. For a swing or pendulum, maximum displacement occurs when the object is at its highest point on either side of the stable equilibrium position. For example, maximum displacement in a spring occurs when the spring reaches the furthest point of being either stretched or compressed.

The amplitude of a cycle is the maximum displacement of particles during a single period of oscillation, and the greater the amplitude, the greater the energy of the oscillation. When an object reaches maximum displacement, it reverses direction, and, therefore, it comes to a stop for an instant of time. Thus, the speed of movement is slowest at that position, and fastest as it passes back through the position of stable equilibrium. An increase in amplitude brings with it an increase in speed, but this does not lead to a change in the period: the greater the amplitude, the further the oscillating object has to move, and, therefore, it takes just as long to complete a cycle.

Restoring Force

Imagine a spring hanging vertically from a ceiling, one end attached to the ceiling for support and the other free to hang. It would thus be in a position of stable equilibrium: the spring hangs downward with a force equal to its weight, and the ceiling pulls it upward with an equal and opposite force. Suppose, now, that the spring is pulled downward.

A spring is highly elastic, meaning that it can experience a temporary stress and still rebound to its original position; by contrast, some objects (for instance, a piece of clay) respond to deformation with plastic behavior, permanently assuming the shape into which they were deformed. The force that directs the spring back to a position of stable equilibrium—the force, in other words, which must be overcome when the spring is pulled downward—is called a restoring force.

The more the spring is stretched, the greater the amount of restoring force that must be overcome. The same is true if the spring is compressed: once again, the spring is removed from a position of equilibrium, and, once again, the restoring force tends to pull it outward to its "natural" position. Here, the example is a spring, but restoring force can be understood just as easily in terms of a swing: once again, it is the force that tends to return the swing to a position of stable equilibrium. There is, however, one significant difference: the restoring force on a swing is gravity, whereas, in the spring, it is related to the properties of the spring itself.

Elastic Potential Energy

For any solid that has not exceeded the elastic limit—the maximum stress it can endure without experiencing permanent deformation—there is a proportional relationship between force and the distance it can be stretched. This is expressed in the formula $F = ks$, where s is the distance and k is a constant related to the size and composition of the material in question.

The amount of force required to stretch the spring is the same as the force that acts to bring it back to equilibrium— that is, the restoring force. Using the value of force, thus derived, it is possible, by a series of steps, to establish a formula for elastic potential energy. The latter, sometimes called strain potential energy, is the potential energy that a spring or a spring-like object possesses by virtue of its deformation from the state of equilibrium. It is equal to $\frac{1}{2}ks^2$.

POTENTIAL AND KINETIC ENERGY. Potential energy, as its name suggests, involves the potential of something to move across a given interval of space—for example, when a sled is perched at the top of a hill. As it begins moving through that interval, the object will gain kinetic energy. Hence, the elastic potential energy of the spring, when the spring is held at a position of the greatest possible displacement from equilibrium, is at a maximum. Once it is released, and the restoring force begins to move it toward the equilibrium position, potential energy drops and kinetic energy increases. But the spring will not just return to equilibrium and stop: its kinetic energy will cause it to keep going.

In the case of the "swing" model of oscillation, elastic potential energy is not a factor. (Unless, of course, the swing itself were suspended on some sort of spring, in which case the object will oscillate in two directions at once.) Nonetheless, all systems of motion involve potential and kinetic energy. When the swing is at a position of maximum displacement, its potential energy is at a maximum as well. Then, as it moves toward the position of stable equilibrium, it loses potential energy and gains kinetic energy. Upon passing through the stable equilib-

A bungee jumper helps illustrate a real-world example of oscillation. *(Eye Ubiquitous/Corbis. Reproduced by permission.)*

rium position, kinetic energy again decreases, while potential energy increases. The sum of the two forms of energy is always the same, but the greater the amplitude, the greater the value of this sum.

REAL-LIFE APPLICATIONS

SPRINGS AND DAMPING

Elastic potential energy relates primarily to springs, but springs are a major part of everyday life. They can be found in everything from the shock-absorber assembly of a motor vehicle to the supports of a trampoline fabric, and in both cases, springs blunt the force of impact.

If one were to jump on a piece of trampoline fabric stretched across an ordinary table—one with no springs—the experience would not be much fun, because there would be little bounce. On the other hand, the elastic potential energy of the trampoline's springs ensures that anyone of normal weight who jumps on the trampoline is liable to bounce some distance into the air. As a person's body comes down onto the trampoline fabric, this stretches the fabric (itself highly elastic) and, hence, the springs. Pulled from a position of equilibrium, the springs acquire elastic potential energy, and this energy makes possible the upward bounce.

As a car goes over a bump, the spring in its shock-absorber assembly is compressed, but the elastic potential energy of the spring immediately forces it back to a position of equilibrium, thus ensuring that the bump is not felt throughout the entire vehicle. However, springs alone would make for a bouncy ride; hence, a modern vehicle also has shock absorbers. The shock absorber, a cylinder in which a piston pushes down on a quantity of oil, acts as a damper—that is, an inhibitor of the springs' oscillation.

SIMPLE HARMONIC MOTION AND DAMPING. Simple harmonic motion occurs when a particle or object moves back and forth within a stable equilibrium position under the influence of a restoring force proportional to its displacement. In an ideal situation, where friction played no part, an object would continue to oscillate indefinitely.

Of course, objects in the real world do not experience perpetual oscillation; instead, most oscillating particles are subject to damping, or the dissipation of energy, primarily as a result of friction. In the earlier illustration of the spring suspended from a ceiling, if the string is pulled to a position of maximum displacement and then released, it will, of course, behave dramatically at first. Over time, however, its movements will become slower and slower, because of the damping effect of frictional forces.

HOW DAMPING WORKS. When the spring is first released, most likely it will fly upward with so much kinetic energy that it will, quite literally, bounce off the ceiling. But with each transit within the position of equilibrium, the friction produced by contact between the metal spring and the air, and by contact between molecules within the spring itself, will gradually reduce the energy that gives it movement. In time, it will come to a stop.

If the damping effect is small, the amplitude will gradually decrease, as the object continues to oscillate, until eventually oscillation ceases. On the other hand, the object may be "overdamped," such that it completes only a few cycles before ceasing to oscillate altogether. In the spring illustration, overdamping would occur if one were to grab the spring on a downward cycle, then slowly let it go, such that it no longer bounced.

There is a type of damping less forceful than overdamping, but not so gradual as the slow dissipation of energy due to frictional forces alone. This is called critical damping. In a critically damped oscillator, the oscillating material is made to return to equilibrium as quickly as possible without oscillating. An example of a critically damped oscillator is the shock-absorber assembly described earlier.

Even without its shock absorbers, the springs in a car would be subject to some degree of damping that would eventually bring a halt to their oscillation; but because this damping is of a very gradual nature, their tendency is to continue oscillating more or less evenly. Over time, of course, the friction in the springs would wear down their energy and bring an end to their oscillation, but by then, the car would most likely have hit another bump. Therefore, it makes sense to apply critical damping to the oscillation of the springs by using shock absorbers.

Bungee Cords and Rubber Bands

Many objects in daily life oscillate in a spring-like way, yet people do not commonly associate them with springs. For example, a rubber band, which behaves very much like a spring, possesses high elastic potential energy. It will oscillate when stretched from a position of stable equilibrium.

Rubber is composed of long, thin molecules called polymers, which are arranged side by side. The chemical bonds between the atoms in a polymer are flexible and tend to rotate, producing kinks and loops along the length of the molecule. The super-elastic polymers in rubber are called elastomers, and when a piece of rubber is pulled, the kinks and loops in the elastomers straighten.

The structure of rubber gives it a high degree of elastic potential energy, and in order to stretch rubber to maximum displacement, there is a powerful restoring force that must be overcome. This can be illustrated if a rubber band is attached to a ceiling, like the spring in the earlier example, and allowed to hang downward. If it is pulled down and released, it will behave much as the spring did.

The oscillation of a rubber band will be even more appreciable if a weight is attached to the "free" end—that is, the end hanging downward. This is equivalent, on a small scale, to a bungee jumper attached to a cord. The type of cord used for bungee jumping is highly elastic; otherwise, the sport would be even more dangerous than it already is. Because of the cord's elasticity, when the bungee jumper "reaches the end of his rope," he bounces back up. At a certain point, he begins to fall again, then bounces back up, and so on, oscillating until he reaches the point of stable equilibrium.

The Pendulum

As noted earlier, a pendulum operates in much the same way as a swing; the difference between them is primarily one of purpose. A swing exists to give pleasure to a child, or a certain bittersweet pleasure to an adult reliving a childhood experience. A pendulum, on the other hand, is not for play; it performs the function of providing a reading, or measurement.

One type of pendulum is a metronome, which registers the tempo or speed of music. Housed in a hollow box shaped like a pyramid, a metronome consists of a pendulum attached to a sliding weight, with a fixed weight attached to the bottom end of the pendulum. It includes a number scale indicating the number of oscillations per minute, and by moving the upper weight, one can change the beat to be indicated.

ZHANG HENG'S SEISMOSCOPE. Metronomes were developed in the early nineteenth century, but, by then, the concept of a pendulum was already old. In the second century A.D., Chinese mathematician and astronomer Zhang Heng (78-139) used a pendulum to develop the world's first seismoscope, an instrument for measuring motion on Earth's surface as a result of earthquakes.

Zhang Heng's seismoscope, which he unveiled in 132 A.D., consisted of a cylinder surrounded by bronze dragons with frogs (also made of bronze) beneath. When the earth shook, a ball would drop from a dragon's mouth into that of a frog, making a noise. The number of balls released, and the direction in which they fell, indicated the magnitude and location of the seismic disruption.

CLOCKS, SCIENTIFIC INSTRUMENTS, AND "FAX MACHINE". In 718 A.D., during a period of intellectual flowering that attended the early T'ang Dynasty (618-907), a Buddhist monk named I-hsing and a military engineer named Liang Ling-tsan built an astronomical clock using a pendulum. Many clocks today—for example, the stately and imposing "grandfather clock" found in some homes—likewise, use a pendulum to mark time.

Physicists of the early modern era used pendula (the plural of pendulum) for a number of interesting purposes, including calculations regarding gravitational force. Experiments with pendula by Galileo Galilei (1564-1642) led to the creation of the mechanical pendulum clock—the grandfather clock, that is—by distinguished Dutch physicist and astronomer Christiaan Huygens (1629-1695).

In the nineteenth century, A Scottish inventor named Alexander Bain (1810-1877) even used a pendulum to create the first "fax machine." Using matching pendulum transmitters and receivers that sent and received electrical

KEY TERMS

AMPLITUDE: The maximum displacement of particles from their normal position during a single period of oscillation.

CYCLE: One full repetition of oscillation. In a single cycle, the oscillating particle moves from a certain point in a certain direction, then switches direction and moves back to the original point. Typically, this is from the position of stable equilibrium to maximum displacement and back again to the stable equilibrium position.

DAMPING: The dissipation of energy during oscillation, which prevents an object from continuing in simple harmonic motion and will eventually force it to stop oscillating altogether. Damping is usually caused by friction.

ELASTIC POTENTIAL ENERGY: The potential energy that a spring or a spring-like object possesses by virtue of its deformation from the state of equilibrium. Sometimes called strain potential energy, it is equal to $½KS^2$, WHERE S is the distance stretched and k is a figure related to the size and composition of the material in question.

EQUILIBRIUM: A state in which the vector sum for all lines of force on an object is equal to zero.

FREQUENCY: For a particle experiencing oscillation, frequency is the number of cycles that take place during one second. Frequency is measured in Hertz.

FRICTION: The force that resists motion when the surface of one object comes into contact with the surface of another.

HARMONIC MOTION: The repeated movement of a particle within a position of equilibrium, or balance.

HERTZ: A unit for measuring frequency. The number of Hertz is the number of cycles per second.

KINETIC ENERGY: The energy that an object possesses due to its motion, as with a sled, when sliding down a hill. This is contrasted with potential energy.

MAXIMUM DISPLACEMENT: For an object in oscillation, maximum displacement is the furthest point from stable equilibrium. Since stable equilibrium is in the middle of a cycle, there are two points of maximum displacement. For a swing or pendulum, this occurs when the object is at its highest point on either side of the stable equilibrium position. Maximum displace-

impulses, he created a crude device that, at the time, seemed to have little practical purpose. In fact, Bain's "fax machine," invented in 1840, was more than a century ahead of its time.

THE FOUCAULT PENDULUM. By far the most important experiments with pendula during the nineteenth century, however, were those of the French physicist Jean Bernard Leon Foucault (1819-1868). Swinging a heavy iron ball from a wire more than 200 ft (61 m) in length, he was able to demonstrate that Earth rotates on its axis.

Foucault conducted his famous demonstration in the Panthéon, a large domed building in Paris named after the ancient Pantheon of Rome. He arranged to have sand placed on the floor of the Panthéon, and placed a pin on the bottom of the iron ball, so that it would mark the sand as the pendulum moved. A pendulum in oscillation maintains its orientation, yet the Foucault pen-

KEY TERMS CONTINUED

ment in a spring occurs when the spring is either stretched or compressed as far as it will go.

OSCILLATION: A type of harmonic motion, typically periodic, in one or more dimensions.

PERIOD: The amount of time required for one cycle in oscillating motion—for instance, from a position of maximum displacement to one of stable equilibrium, and, once again, to maximum displacement.

PERIODIC MOTION: Motion that is repeated at regular intervals. These intervals are known as periods.

POTENTIAL ENERGY: The energy that an object possesses due to its position, as for instance, with a sled at the top of a hill. This is contrasted with kinetic energy.

RESTORING FORCE: A force that directs an object back to a position of stable equilibrium. An example is the resistance of a spring, when it is extended.

SIMPLE HARMONIC MOTION: Harmonic motion, in which a particle moves back and forth about a stable equilibrium position under the influence of a restoring force proportional to its displacement. Simple harmonic motion is, in fact, an ideal situation; most types of oscillation are subject to some form of damping.

STABLE EQUILIBRIUM: A type of equilibrium in which, if an object were disturbed, it would tend to return to its original position. For an object in oscillation, stable equilibrium is in the middle of a cycle, between two points of maximum displacement.

VECTOR: A quantity that possesses both magnitude and direction.

VECTOR SUM: A calculation that yields the net result of all the vectors applied in a particular situation. Because direction is involved, it is necessary when calculating the vector sum of forces on an object (as, for instance, when determining whether or not it is in a state of equilibrium), to assign a positive value to forces in one direction, and a negative value to forces in the opposite direction. If the object is in equilibrium, these forces will cancel one another out.

dulum (as it came to be called) seemed to be shifting continually toward the right, as indicated by the marks in the sand.

The confusion related to reference point: since Earth's rotation is not something that can be perceived with the senses, it was natural to assume that the pendulum itself was changing orientation—or rather, that only the pendulum was moving. In fact, the path of Foucault's pendulum did not vary nearly as much as it seemed. Earth itself was moving beneath the pendulum, providing an additional force which caused the pendulum's plane of oscillation to rotate.

WHERE TO LEARN MORE

Brynie, Faith Hickman. *Six-Minute Science Experiments.* Illustrated by Kim Whittingham. New York: Sterling Publishing Company, 1996.

Ehrlich, Robert. *Turning the World Inside Out, and 174 Other Simple Physics Demonstrations.* Princeton, N.J.: Princeton University Press, 1990.

"*Foucault Pendulum*" *Smithsonian Institution FAQs* (Web site). <http://www.si.edu/resource/faq/nmah/pendulum.html> (April 23, 2001).

Kruszelnicki, Karl S. *The Foucault Pendulum* (Web site). <http://www.abc.net.au/surf/pendulum/pendulum.html> (April 23, 2001).

Schaefer, Lola M. *Back and Forth.* Edited by Gail Saunders-Smith; P. W. Hammer, consultant. Mankato, MN: Pebble Books, 2000.

Shirley, Jean. *Galileo.* Illustrated by Raymond Renard. St. Louis: McGraw-Hill, 1967.

Suplee, Curt. *Everyday Science Explained.* Washington, D.C.: National Geographic Society, 1996.

Topp, Patricia. *This Strange Quantum World and You.* Nevada City, CA: Blue Dolphin, 1999.

Zubrowski, Bernie. *Making Waves: Finding Out About Rhythmic Motion.* Illustrated by Roy Doty. New York: Morrow Junior Books, 1994.

FREQUENCY

CONCEPT

Everywhere in daily life, there are frequencies of sound and electromagnetic waves, constantly changing and creating the features of the visible and audible world familiar to everyone. Some aspects of frequency can only be perceived indirectly, yet people are conscious of them without even thinking about it: a favorite radio station, for instance, may have a frequency of 99.7 MHz, and fans of that station knows that every time they turn the FM dial to that position, the station's signal will be there. Of course, people cannot "hear" radio and television frequencies—part of the electromagnetic spectrum—but the evidence for them is everywhere. Similarly, people are not conscious, in any direct sense, of frequencies in sound and light—yet without differences in frequency, there could be no speech or music, nor would there be any variations of color.

HOW IT WORKS

Harmonic Motion and Energy

In order to understand frequency, it is first necessary to comprehend two related varieties of movement: oscillation and wave motion. Both are examples of a broader category, periodic motion: movement that is repeated at regular intervals called periods. Oscillation and wave motion are also examples of harmonic motion, or the repeated movement of a particle about a position of equilibrium, or balance.

KINETIC AND POTENTIAL ENERGY. In harmonic motion, and in some types of periodic motion, there is a continual conversion of energy from one form to another. On the one hand is potential energy, or the energy of an object due to its position and, hence, its potential for movement. On the other hand, there is kinetic energy, the energy of movement itself.

Potential-kinetic conversions take place constantly in daily life: any time an object is at a distance from a position of stable equilibrium, and some force (for instance, gravity) is capable of moving it to that position, it possesses potential energy. Once it begins to move toward that equilibrium position, it loses potential energy and gains kinetic energy. Likewise, a wave at its crest has potential energy, and gains kinetic energy as it moves toward its trough. Similarly, an oscillating object that is as far as possible from the stable-equilibrium position has enormous potential energy, which dissipates as it begins to move toward stable equilibrium.

VIBRATION. Though many examples of periodic and harmonic motion can be found in daily life, the terms themselves are certainly not part of everyday experience. On the other hand, everyone knows what "vibration" means: to move back and forth in place. Oscillation, discussed in more detail below, is simply a more scientific term for vibration; and while waves are not themselves merely vibrations, they involve—and may produce—vibrations. This, in fact, is how the human ear hears: by interpreting vibrations resulting from sound waves.

Indeed, the entire world is in a state of vibration, though people seldom perceive this movement—except, perhaps, in dramatic situations such as earthquakes, when the vibrations of plates beneath Earth's surface become too force-

GRANDFATHER CLOCKS ARE ONE OF THE BEST-KNOWN VARIETIES OF A PENDULUM. *(Photograph by Peter Harholdt/Corbis. Reproduced by permission.)*

ful to ignore. All matter vibrates at the molecular level, and every object possesses what is called a natural frequency, which depends on its size, shape, and composition. This explains how a singer can shatter a glass by hitting a certain note, which does not happen because the singer's voice has reached a particularly high pitch; rather, it is a matter of attaining the natural frequency of the glass. As a result, all the energy in the sound of the singer's voice is transferred to the glass, and it shatters.

OSCILLATION

Oscillation is a type of harmonic motion, typically periodic, in one or more dimensions. There are two basic types of oscillation: that of a swing or pendulum and that of a spring. In each case, an object is disturbed from a position of stable equilibrium, and, as a result, it continues to move back and forth around that stable equilibrium position. If a spring is pulled from stable equilibrium, it will generally oscillate along a straight path; a swing, on the other hand, will oscillate along an arc.

In oscillation, whether the oscillator be spring-like or swing-like, there is always a cycle in which the oscillating particle moves from a certain point in a certain direction, then reverses direction and returns to the original point. Usually a cycle is viewed as the movement from a position of stable equilibrium to one of maximum displacement, or the furthest possible point from stable equilibrium. Because stable equilibrium is directly in the middle of a cycle, there are two points of maximum displacement: on a swing, this occurs when the object is at its highest point on either side of the stable equilibrium position, and on a spring, maximum displacement occurs when the spring is either stretched or compressed as far as it will go.

WAVE MOTION

Wave motion is a type of harmonic motion that carries energy from one place to another without actually moving any matter. While oscillation involves the movement of "an object," whether it be a pendulum, a stretched rubber band, or some other type of matter, a wave may or may not involve matter. Example of a wave made out of matter—that is, a mechanical wave—is a wave on the ocean, or a sound wave, in which energy vibrates through a medium such as air. Even in the case of the mechanical wave, however, the matter does not experience any net displacement from its original position. (Water molecules do rotate as a result of wave motion, but they end up where they began.)

There are waves that do not follow regular, repeated patterns; however, within the context of frequency, our principal concern is with periodic waves, or waves that follow one another in regular succession. Examples of periodic waves include ocean waves, sound waves, and electromagnetic waves.

Periodic waves may be further divided into transverse and longitudinal waves. A transverse wave is the shape that most people imagine when they think of waves: a regular up-and-down pattern (called "sinusoidal" in mathematical terms) in which the vibration or motion is perpendicular to the direction the wave is moving.

A longitudinal wave is one in which the movement of vibration is in the same direction as the wave itself. Though these are a little harder to picture, longitudinal waves can be visualized as a series of concentric circles emanating from a single point. Sound waves are longitudinal: thus when someone speaks, waves of sound vibrations radiate out in all directions.

Amplitude

There are certain properties of waves, such as wavelength, or the distance between waves, that are not properties of oscillation. However, both types of motion can be described in terms of amplitude, period, and frequency. The first of these is not related to frequency in any mathematical sense; nonetheless, where sound waves are concerned, both amplitude and frequency play a significant role in what people hear.

Though waves and oscillators share the properties of amplitude, period, and frequency, the definitions of these differ slightly depending on whether one is discussing wave motion or oscillation. Amplitude, generally speaking, is the value of maximum displacement from an average value or position—or, in simpler terms, amplitude is "size." For an object experiencing oscillation, it is the value of the object's maximum displacement from a position of stable equilibrium during a single period. It is thus the "size" of the oscillation.

In the case of wave motion, amplitude is also the "size" of a wave, but the precise definition varies, depending on whether the wave in question is transverse or longitudinal. In the first instance, amplitude is the distance from either the crest or the trough to the average position between them. For a sound wave, which is longitudinal, amplitude is the maximum value of the pressure change between waves.

Period and Frequency

Unlike amplitude, period is directly related to frequency. For a transverse wave, a period is the amount of time required to complete one full cycle of the wave, from trough to crest and back to trough. In a longitudinal wave, a period is the interval between waves. With an oscillator, a period is the amount of time it takes to complete one cycle. The value of a period is usually expressed in seconds.

Frequency in oscillation is the number of cycles per second, and in wave motion, it is the number of waves that pass through a given point per second. These cycles per second are called Hertz (Hz) in honor of nineteenth-century German physicist Heinrich Rudolf Hertz (1857-1894), who greatly advanced understanding of electromagnetic wave behavior during his short career.

Middle C—which is at the middle of a piano keyboard—is the starting point of a basic musical scale. It is called the fundamental frequency, or the first harmonic. *(Photograph by Francoise Gervais/Corbis. Reproduced by permission.)*

If something has a frequency of 100 Hz, this means that 100 waves are passing through a given point during the interval of one second, or that an oscillator is completing 100 cycles in a second. Higher frequencies are expressed in terms of kilohertz (kHz; 10^3 or 1,000 cycles per second); megahertz (MHz; 10^6 or 1 million cycles per second); and gigahertz (GHz; 10^9 or 1 billion cycles per second.).

A clear mathematical relationship exists between period, symbolized by *T,* and frequency (*f*): each is the inverse of the other. Hence,

$$T = \frac{1}{f},$$

and

$$f = \frac{1}{T}.$$

If an object in harmonic motion has a frequency of 50 Hz, its period is 1/50 of a second (0.02 sec). Or, if it has a period of 1/20,000 of a second (0.00005 sec), that means it has a frequency of 20,000 Hz.

REAL-LIFE APPLICATIONS

Grandfather Clocks and Metronomes

One of the best-known varieties of pendulum (plural, pendula) is a grandfather clock. Its invention was an indirect result of experiments with pendula by Galileo Galilei (1564-1642), work that influenced Dutch physicist and astronomer Christiaan Huygens (1629-1695) in the creation of the mechanical pendulum clock—or grandfather clock, as it is commonly known.

The frequency of a pendulum, a swing-like oscillator, is the number of "swings" per minute. Its frequency is proportional to the square root of the downward acceleration due to gravity (32 ft or 9.8 m/sec^2) divided by the length of the pendulum. This means that by adjusting the length of the pendulum on the clock, one can change its frequency: if the pendulum length is shortened, the clock will run faster, and if it is lengthened, the clock will run more slowly.

Another variety of pendulum, this one dating to the early nineteenth century, is a metronome, an instrument that registers the tempo or speed of music. Consisting of a pendulum attached to a sliding weight, with a fixed weight attached to the bottom end of the pendulum, a metronome includes a number scale indicating the frequency—that is, the number of oscillations per minute. By moving the upper weight, one can speed up or slow down the beat.

Harmonics

As noted earlier, the volume of any sound is related to the amplitude of the sound waves. Frequency, on the other hand, determines the pitch or tone. Though there is no direct correlation between intensity and frequency, in order for a person to hear a very low-frequency sound, it must be above a certain decibel level.

The range of audibility for the human ear is from 20 Hz to 20,000 Hz. The optimal range for hearing, however, is between 3,000 and 4,000 Hz. This places the piano, whose 88 keys range from 27 Hz to 4,186 Hz, well within the range of human audibility. Many animals have a much wider range: bats, whales, and dolphins can hear sounds at a frequency up to 150,000 Hz. But humans have something that few animals can appreciate: music, a realm in which frequency changes are essential.

Each note has its own frequency: middle C, for instance, is 264 Hz. But in order to produce what people understand as music—that is, pleasing combinations of notes—it is necessary to employ principles of harmonics, which express the relationships between notes. These mathematical relations between musical notes are among the most intriguing aspects of the connection between art and science.

It is no wonder, perhaps, that the great Greek mathematician Pythagoras (c. 580-500 B.C.) believed that there was something spiritual or mystical in the connection between mathematics and music. Pythagoras had no concept of frequency, of course, but he did recognize that there were certain numerical relationships between the lengths of strings, and that the production of harmonious music depended on these ratios.

RATIOS OF FREQUENCY AND PLEASING TONES. Middle C—located,, appropriately enough, in the middle of a piano keyboard—is the starting point of a basic musical scale. It is called the fundamental frequency, or the first harmonic. The second harmonic, one octave above middle C, has a frequency of 528 Hz, exactly twice that of the first harmonic; and the third harmonic (two octaves above middle C) has a frequency of 792 cycles, or three times that of middle C. So it goes, up the scale.

As it turns out, the groups of notes that people consider harmonious just happen to involve specific whole-number ratios. In one of those curious interrelations of music and math that would have delighted Pythagoras, the smaller the numbers involved in the ratios, the more pleasing the tone to the human psyche.

An example of a pleasing interval within an octave is a fifth, so named because it spans five notes that are a whole step apart. The C Major scale is easiest to comprehend in this regard, because it does not require reference to the "black keys," which are a half-step above or below the "white keys." Thus, the major fifth in the C-Major scale is C, D, E, F, G. It so happens that the

KEY TERMS

AMPLITUDE: For an object oscillation, amplitude is the value of the object's maximum displacement from a position of stable equilibrium during a single period. In a transverse wave, amplitude is the distance from either the crest or the trough to the average position between them. For a sound wave, the best-known example of a longitudinal wave, amplitude is the maximum value of the pressure change between waves.

CYCLE: In oscillation, a cycle occurs when the oscillating particle moves from a certain point in a certain direction, then switches direction and moves back to the original point. Typically, this is from the position of stable equilibrium to maximum displacement and back again to the stable equilibrium position.

FREQUENCY: For a particle experiencing oscillation, frequency is the number of cycles that take place during one second. In wave motion, frequency is the number of waves passing through a given point during the interval of one second. In either case, frequency is measured in Hertz. Period (T) is the mathematical inverse of frequency (f) hence $f=1/T$.

HARMONIC MOTION: The repeated movement of a particle about a position of equilibrium, or balance.

HERTZ: A unit for measuring frequency, named after nineteenth-century German physicist Heinrich Rudolf Hertz (1857-1894). Higher frequencies are expressed in terms of kilohertz (kHz; 10^3 or 1,000 cycles per second); megahertz (MHz; 10^6 or 1 million cycles per second); and gigahertz (GHz; 10^9 or 1 billion cycles per second.)

KINETIC ENERGY: The energy that an object possesses due to its motion, as with a sled when sliding down a hill. This is contrasted with potential energy.

LONGITUDINAL WAVE: A wave in which the movement of vibration is in the same direction as the wave itself. This is contrasted to a transverse wave.

MAXIMUM DISPLACEMENT: For an object in oscillation, maximum displacement is the farthest point from stable equilibrium.

OSCILLATION: A type of harmonic motion, typically periodic, in one or more dimensions.

PERIOD: In oscillation, a period is the amount of time required for one cycle. For a transverse wave, a period is the amount of time required to complete one full cycle of the wave, from trough to crest and back to trough. In a longitudinal wave, a period is the interval between waves. Frequency is the mathematical inverse of period (T): hence, $T=1/f$.

PERIODIC MOTION: Motion that is repeated at regular intervals. These intervals are known as periods.

KEY TERMS CONTINUED

POTENTIAL ENERGY: The energy that an object possesses due to its position, as, for instance, with a sled at the top of a hill. This is contrasted with kinetic energy.

STABLE EQUILIBRIUM: A position in which, if an object were disturbed, it would tend to return to its original position. For an object in oscillation, stable equilibrium is in the middle of a cycle, between two points of maximum displacement.

TRANSVERSE WAVE: A wave in which the vibration or motion is perpendicular to the direction in which the wave is moving. This is contrasted to a longitudinal wave.

WAVE MOTION: A type of harmonic motion that carries energy from one place to another without actually moving any matter.

ratio in frequency between middle C and G (396 Hz) is 2:3.

Less melodious, but still certainly tolerable, is an interval known as a third. Three steps up from middle C is E, with a frequency of 330 Hz, yielding a ratio involving higher numbers than that of a fifth—4:5. Again, the higher the numbers involved in the ratio, the less appealing the sound is to the human ear: the combination E-F, with a ratio of 15:16, sounds positively grating.

The Electromagnetic Spectrum

Everyone who has vision is aware of sunlight, but, in fact, the portion of the electromagnetic spectrum that people perceive is only a small part of it. The frequency range of visible light is from $4.3 \cdot 10^{14}$ Hz to $7.5 \cdot 10^{14}$ Hz—in other words, from 430 to 750 trillion Hertz. Two things should be obvious about these numbers: that both the range and the frequencies are extremely high. Yet, the values for visible light are small compared to the higher reaches of the spectrum, and the range is also comparatively small.

Each of the colors has a frequency, and the value grows higher from red to orange, and so on through yellow, green, blue, indigo, and violet. Beyond violet is ultraviolet light, which human eyes cannot see. At an even higher frequency are x rays, which occupy a broad band extending almost to 10^{20} Hz—in other words, 1 followed by 20 zeroes. Higher still is the very broad range of gamma rays, reaching to frequencies as high as 10^{25}. The latter value is equal to 10 trillion trillion.

Obviously, these ultra-ultra high-frequency waves must be very small, and they are: the higher gamma rays have a wavelength of around 10^{-15} meters (0.000000000000001 m). For frequencies lower than those of visible light, the wavelengths get larger, but for a wide range of the electromagnetic spectrum, the wavelengths are still much too small to be seen, even if they were visible. Such is the case with infrared light, or the relatively lower-frequency millimeter waves.

Only at the low end of the spectrum, with frequencies below about 10^{10} Hz—still an incredibly large number—do wavelengths become the size of everyday objects. The center of the microwave range within the spectrum, for instance, has a wavelength of about 3.28 ft (1 m). At this end of the spectrum—which includes television and radar (both examples of microwaves), short-wave radio, and long-wave radio—there are numerous segments devoted to various types of communication.

RADIO AND MICROWAVE FREQUENCIES. The divisions of these sections of the electromagnetic spectrum are arbitrary and manmade, but in the United States—where they are administered by the Federal Communications Commission (FCC)—they have the force of law. When AM (amplitude modulation) radio first came into widespread use in the early

1920s—Congress assigned AM stations the frequency range that they now occupy: 535 kHz to 1.7 MHz.

A few decades after the establishment of the FCC in 1927, new forms of electronic communication came into being, and these too were assigned frequencies—sometimes in ways that were apparently haphazard. Today, television stations 2-6 are in the 54-88 MHz range, while stations 7-13 occupy the region from 174-220 MHz. In between is the 88 to 108 MHz band, assigned to FM radio. Likewise, short-wave radio (5.9 to 26.1 MHz) and citizens' band or CB radio (26.96 to 27.41 MHz) occupy positions between AM and FM.

In fact, there are a huge variety of frequency ranges accorded to all manner of other communication technologies. Garage-door openers and alarm systems have their place at around 40 MHz. Much, much higher than these—higher, in fact, than TV broadcasts—is the band allotted to deep-space radio communications: 2,290 to 2,300 MHz. Cell phones have their own realm, of course, as do cordless phones; but so too do radio controlled cars (75 MHz) and even baby monitors (49 MHz).

WHERE TO LEARN MORE

Beiser, Arthur. Physics, 5th ed. Reading, MA: Addison-Wesley, 1991.

Allocation of Radio Spectrum in the United States (Web site). <http://members.aol.com/jneuhaus/fccindex/spectrum.html> (April 25, 2001).

DiSpezio, Michael and Catherine Leary. *Awesome Experiments in Light and Sound.* New York: Sterling Juvenile, 2001.

Electromagnetic Spectrum (Web site). <http://www.jsc.mil/images/speccht.jpg> (April 25, 2001).

"How the Radio Spectrum Works." How Stuff Works (Web site). <http://www.howstuffworks.com/radio~spectrum.html> (April 25, 2001).

Internet Resources for Sound and Light (Web site). <http://electro.sau.edu/SLResources.html> (April 25, 2001).

"NIST Time and Frequency Division." NIST: National Institute of Standards and Technology (Web site). <http://www.boulder.nist.gov/timefreq/> (April 25, 2001).

Parker, Steve. *Light and Sound.* Austin, TX: Raintree Steck-Vaughn, 2000.

Physics Tutorial System: Sound Waves Modules (Web site). <http://csgrad.cs.vt.edu/~chin/chin_sound.html> (April 25, 2001).

"Radio Electronics Pages" ePanorama.net (Web site). <http://www.epanorama.net/radio.html> (April 25, 2001).

RESONANCE

CONCEPT

Though people seldom witness it directly, the entire world is in a state of motion, and where solid objects are concerned, this motion is manifested as vibration. When the vibrations produced by one object come into alignment with those of another, this is called resonance. The power of resonance can be as gentle as an adult pushing a child on a swing, or as ferocious as the force that toppled what was once the world's third-longest suspension bridge. Resonance helps to explain all manner of familiar events, from the feedback produced by an electric guitar to the cooking of food in a microwave oven.

HOW IT WORKS

Vibration of Molecules

The possibility of resonance always exists wherever there is periodic motion, movement that is repeated at regular intervals called periods, and/or harmonic motion, the repeated movement of a particle about a position of equilibrium or balance. Many examples of resonance involve large objects: a glass, a child on a swing, a bridge. But resonance also takes place at a level invisible to the human eye using even the most powerful optical microscope.

All molecules exert a certain electromagnetic attraction toward each other, and generally speaking, the less the attraction between molecules, the greater their motion relative to one another. This, in turn, helps define the object in relation to its particular phase of matter.

A substance in which molecules move at high speeds, and therefore hardly attract one another at all, is called a gas. Liquids are materials in which the rate of motion, and hence of intermolecular attraction, is moderate. In a solid, on the other hand, there is little relative motion, and therefore molecules exert enormous attractive forces. Instead of moving in relation to one another, the molecules that make up a solid tend to vibrate in place.

Due to the high rate of motion in gas molecules, gases possess enormous internal kinetic energy. The internal energy of solids and liquids is much less than in gases, yet, as we shall see, the use of resonance to transfer energy to these objects can yield powerful results.

Oscillation

In colloquial terms, oscillation is the same as vibration, but, in more scientific terms, oscillation can be identified as a type of harmonic motion, typically periodic, in one or more dimensions. All things that oscillate do so either along a more or less straight path, like that of a spring pulled from a position of stable equilibrium; or they oscillate along an arc, like a swing or pendulum.

In the case of the swing or pendulum, stable equilibrium is the point at which the object is hanging straight downward—that is, the position to which gravitation force would take it if no other net forces were acting on the object. For a spring, stable equilibrium lies somewhere between the point at which the spring is stretched to its maximum length and the point at which it is subjected to maximum compression without permanent deformation.

CYCLES AND FREQUENCY. A cycle of oscillation involves movement from a certain point in a certain direction, then a reversal of direction and a return to the original point. It is simplest to treat a cycle as the movement from a position of stable equilibrium to one of maximum displacement, or the furthest possible point from stable equilibrium.

The amount of time it takes to complete one cycle is called a period, and the number of cycles in one second is the frequency of the oscillation. Frequency is measured in Hertz. Named after nineteenth-century German physicist Heinrich Rudolf Hertz (1857-1894), a single Hertz (Hz)—the term is both singular and plural—is equal to one cycle per second.

AMPLITUDE AND ENERGY. The amplitude of a cycle is the maximum displacement of particles during a single period of oscillation. When an oscillator is at maximum displacement, its potential energy is at a maximum as well. From there, it begins moving toward the position of stable equilibrium, and as it does so, it loses potential energy and gains kinetic energy. Once it reaches the stable equilibrium position, kinetic energy is at a maximum and potential energy at a minimum.

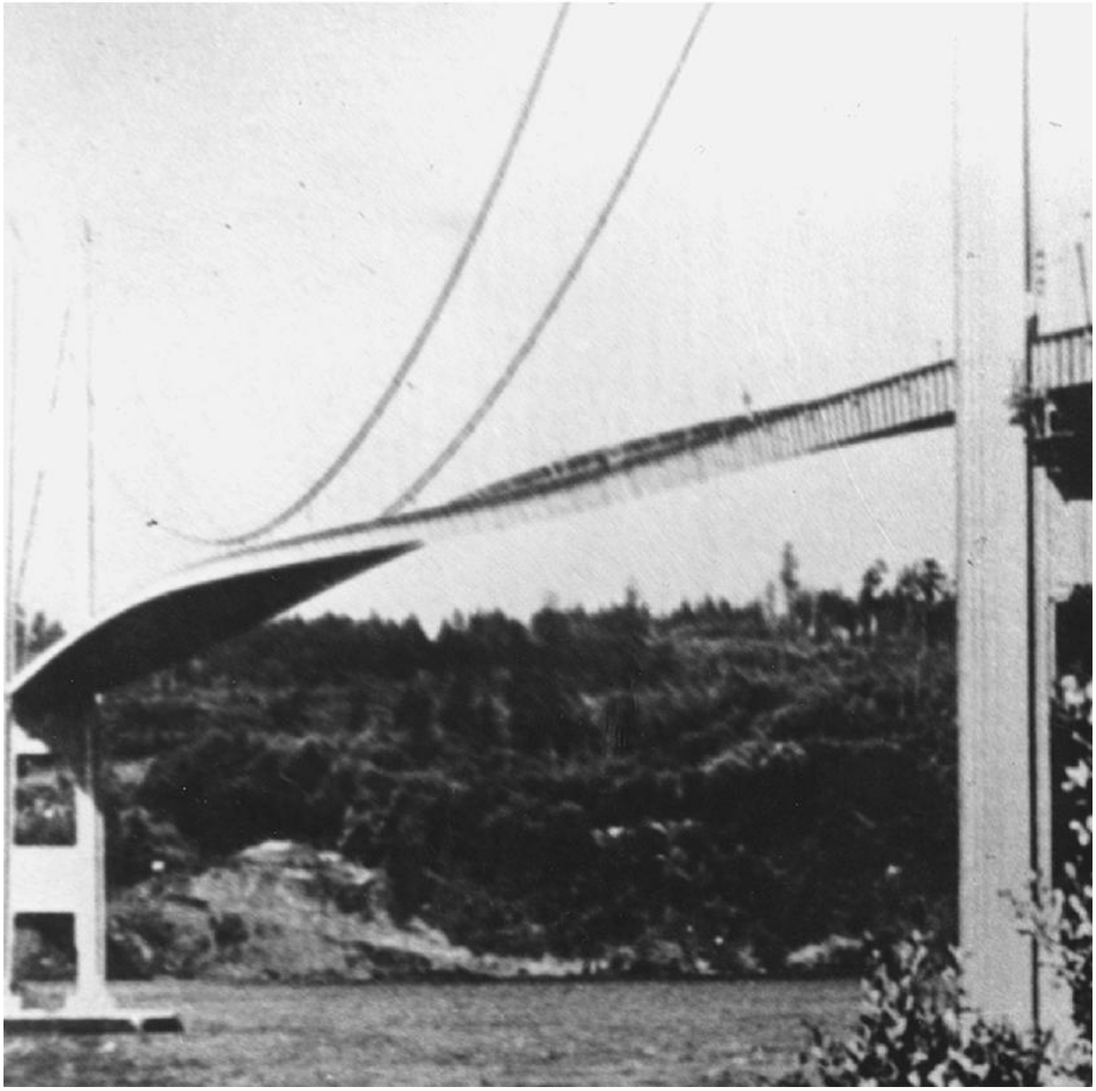
As the oscillating object passes through the position of stable equilibrium, kinetic energy begins to decrease and potential energy increases. By the time it has reached maximum displacement again—this time on the other side of the stable equilibrium position—potential energy is once again at a maximum.

OSCILLATION IN WAVE MOTION. The particles in a mechanical wave (a wave that moves through a material medium) have potential energy at the crest and trough, and gain kinetic energy as they move between these points. This is just one of many ways in which wave motion can be compared to oscillation. There is one critical difference between oscillation and wave motion: whereas oscillation involves no net movement, but merely movement in place, the harmonic motion of waves carries energy from one place to another. Nonetheless, the analogies than can be made between waves and oscillations are many, and understandably so: oscillation, after all, is an aspect of wave motion.

A periodic wave is one in which a uniform series of crests and troughs follow one after the other in regular succession. Two basic types of periodic waves exist, and these are defined by the relationship between the direction of oscillation and the direction of the wave itself. A transverse wave forms a regular up-and-down pattern, in which the oscillation is perpendicular to the direction in which the wave is moving. On the other hand, in a longitudinal wave (of which a sound wave is the best example), oscillation is in the same direction as the wave itself.

A COMMON EXAMPLE OF RESONANCE: A PARENT PUSHES HER CHILD ON A SWING. *(Photograph by Annie Griffiths Belt/Corbis. Reproduced by permission.)*

Again, the wave itself experiences net movement, but within the wave—one of its defining characteristics, as a matter of fact—are oscillations, which (also by definition) experience no net movement. In a transverse wave, which is usually easier to visualize than a longitudinal wave, the oscillation is from the crest to the trough and back again. At the crest or trough, potential energy is at a maximum, while kinetic energy reaches a maximum at the point of equilibrium between crest and trough. In a longitudinal wave, oscillation is a matter of density fluctuations: the greater the value of these fluctuations, the greater the energy in the wave.



THE POWER OF RESONANCE CAN DESTROY A BRIDGE. ON NOVEMBER 7, 1940, THE ACCLAIMED TACOMA NARROWS BRIDGE COLLAPSED DUE TO OVERWHELMING RESONANCE. *(UPI/Corbis-Bettmann. Reproduced by permission.)*

PARAMETERS FOR DESCRIBING HARMONIC MOTION

The maximum value of the pressure change between waves is the amplitude of a longitudinal wave. In fact, waves can be described according to many of the same parameters used for oscillation—frequency, period, amplitude, and so on. The definitions of these terms vary somewhat, depending on whether one is discussing oscillation or wave motion; or, where wave motion is concerned, on whether the subject is a transverse wave or a longitudinal wave.

For the present purposes, however, it is necessary to focus on just a few specifics of harmonic motion. First of all, the type of motion with which we will be concerned is oscillation, and though wave motion will be mentioned, our principal concern is the oscillations within the waves, not the waves themselves. Second, the two parameters of importance in understanding resonance are amplitude and frequency.

RESONANCE AND ENERGY TRANSFER

Resonance can be defined as the condition in which force is applied to an oscillator at the point of maximum amplitude. In this way, the motion of the outside force is perfectly matched to that of the oscillator, making possible a transfer of energy.

As its name suggests, resonance is a matter of one object or force "getting in tune with" another object. One literal example of this involves shattering a wine glass by hitting a musical note that is on the same frequency as the natural frequency of the glass. (Natural frequency depends on the size, shape, and composition of the object in question.) Because the frequencies resonate, or are in sync with one another, maximum energy transfer is possible.

The same can be true of soldiers walking across a bridge, or of winds striking the bridge at a resonant frequency—that is, a frequency that matches that of the bridge. In such situations, a large structure may collapse under a force that would not normally destroy it, but the effects of resonance are not always so dramatic. Sometimes resonance can be a simple matter, like pushing a child in a swing in such a way as to ensure that the child gets maximum enjoyment for the effort expended.

REAL-LIFE APPLICATIONS

A Child on a Swing and a Pendulum in a Museum

Suppose a father is pushing his daughter on a swing, so that she glides back and forth through the air. A swing, as noted earlier, is a classic example of an oscillator. When the child gets in the seat, the swing is in a position of stable equilibrium, but as the father pulls her back before releasing her, she is at maximum displacement.

He releases her, and quickly, potential energy becomes kinetic energy as she swings toward the position of stable equilibrium, then up again on the other side. Now the half-cycle is repeated, only in reverse, as she swings backward toward her father. As she reaches the position from which he first pushed her, he again gives her a little push. This push is essential, if she is to keep going. Without friction, she could keep on swinging forever at the same rate at which she begun. But in the real world, the wearing of the swing's chain against the support along the bar above the swing will eventually bring the swing itself to a halt.

TIMING THE PUSH. Therefore, the father pushes her—but in order for his push to be effective, he must apply force at just the right moment. That right moment is the point of greatest amplitude—the point, that is, at which the father's pushing motion and the motion of the swing are in perfect resonance.

If the father waits until she is already on the downswing before he pushes her, not all the energy of his push will actually be applied to keeping her moving. He will have failed to efficiently add energy to his daughter's movement on the swing. On the other hand, if he pushes her too soon—that is, while she is on the upswing—he will actually take energy away from her movement.

If his purpose were to bring the swing to a stop, then it would make good sense to push her on the upswing, because this would produce a cycle of smaller amplitude and hence less energy. But if the father's purpose is to help his daughter keep swinging, then the time to apply energy is at the position of maximum displacement.

It so happens that this is also the position at which the swing's speed is the slowest. Once it reaches maximum displacement, the swing is about to reverse direction, and, therefore, it stops for a split-second. Once it starts moving again, now in a new direction, both kinetic energy and speed increase until the swing passes through the position of stable equilibrium, where it reaches its highest rate.

THE FOUCAULT PENDULUM. Hanging from a ceiling in Washington, D.C.'s Smithsonian Institution is a pendulum 52 ft (15.85 m) long, at the end of which is an iron ball weighing 240 lb (109 kg). Back and forth it swings, and if one sits and watches it long enough, the pendulum appears to move gradually toward the right. Over the course of 24 hours, in fact, it seems to complete a full circuit, moving back to its original orientation.

There is just one thing wrong with this picture: though the pendulum is shifting direction, this does not nearly account for the total change in orientation. At the same time the pendulum is moving, Earth is rotating beneath it, and it is the viewer's frame of reference that creates the mistaken impression that only the pendulum is rotating. In fact it is oscillating, swinging back and forth from the Smithsonian ceiling, but though it shifts orientation somewhat, the greater component of this shift comes from the movement of the Earth itself.

This particular type of oscillator is known as a Foucault pendulum, after French physicist Jean Bernard Leon Foucault (1819-1868), who in 1851 used just such an instrument to prove that Earth is rotating. Visitors to the Smithsonian, after they get over their initial bewilderment at the fact that the pendulum is not actually rotating, may well have another question: how exactly does the pendulum keep moving?

As indicated earlier, in an ideal situation, a pendulum continues oscillating. But situations on Earth are not ideal: with each swing, the Foucault pendulum loses energy, due to friction from the air through which it moves. In addition, the cable suspending it from the ceiling is also oscillating slightly, and this, too, contributes to energy loss. Therefore, it is necessary to add energy to the pendulum's swing.

Surrounding the cable where it attaches to the ceiling is an electromagnet shaped like a donut, and on either side, near the top of the cable, are two iron collars. An electronic device senses when the pendulum reaches maximum amplitude, switching on the electromagnet, which causes the appropriate collar to give the cable a slight jolt. Because the jolt is delivered at the right moment, the resonance is perfect, and energy is restored to the pendulum.

Resonance in Electricity and Electromagnetic Waves

Resonance is a factor in electromagnetism, and in electromagnetic waves, such as those of light or radio. Though much about electricity tends to be rather abstract, the idea of current is fairly easy to understand, because it is more or less analogous to a water current: hence, the less impedance to flow, the stronger the current. Minimal impedance is achieved when the impressed voltage has a certain resonant frequency.

NUCLEAR MAGNETIC RESONANCE. The term "nuclear magnetic resonance" (NMR) is hardly a household world, but thanks to its usefulness in medicine, MRI—short for magnetic resonance imagining—is certainly a well-known term. In fact, MRI is simply the medical application of NMR. The latter is a process in which a rotating magnetic field is produced, causing the nuclei of certain atoms to absorb energy from the field. It is used in a range of areas, from making nuclear measurements to medical imaging, or MRI. In the NMR process, the nucleus of an atom is forced to wobble like a top, and this speed of wobbling is increased by applying a magnetic force that resonates with the frequency of the wobble.

The principles of NMR were first developed in the late 1930s, and by the early 1970s they had been applied to medicine. Thanks to MRI, physicians can make diagnoses without the patient having to undergo either surgery or x rays. When a patient undergoes MRI, he or she is made to lie down inside a large tube-like chamber. A technician then activates a powerful magnetic field that, depending on its position, resonates with the frequencies of specific body tissues. It is thus possible to isolate specific cells and analyze them independently, a process that would be virtually impossible otherwise without employing highly invasive procedures.

LIGHT AND RADIO WAVES. One example of resonance involving visible and invisible light in the electromagnetic spectrum is resonance fluorescence. Fluorescence itself is a process whereby a material absorbs electromagnetic radiation from one source, then re-emits that radiation on a wavelength longer than that of the illuminating radiation. Among its many applications are the fluorescent lights found in many homes and public buildings. Sometimes the emitted radiation has the same wavelength as the absorbed radiation, and this is called resonance fluorescence. Resonance fluorescence is used in laboratories for analyzing phenomena such as the flow of gases in a wind tunnel.

Though most people do not realize that radio waves are part of the electromagnetic spectrum, radio itself is certainly a part of daily life, and, here again, resonance plays a part. Radio waves are relatively large compared to visible light waves, and still larger in comparison to higher-frequency waves, such as those in ultraviolet light or x rays. Because the wavelength of a radio signal is as large as objects in ordinary experience, there can sometimes be conflict if the size of an antenna does not match properly with a radio wave. When the sizes are compatible, this, too, is an example of resonance.

MICROWAVES. Microwaves occupy a part of the electromagnetic spectrum with higher frequencies than those of radio waves. Examples of microwaves include television signals, radar—and of course the microwave oven, which cooks food without applying external heat. Like many other useful products, the microwave oven ultimately arose from military-industrial research, in this case, during World War II. Introduced for home use in 1955, its popularity grew slowly for the first few decades, but in the 1970s and 1980s, microwave use increased dramatically. Today, most American homes have microwaves ovens.

Of course there will always be types of food that cook better in a conventional oven, but the beauty of a microwave is that it makes possible the quick heating and cooking of foods—all without the drying effect of conventional baking. The basis for the microwave oven is the fact that the molecules in all forms of matter are vibrating. By achieving resonant frequency, the oven adds energy—heat—to food. The oven is not equipped in such a way as to detect the frequency of molecular vibration in all possible substances, however; instead, the microwaves resonant with the frequency of a single item found in nearly all types of food: water.

Emitted from a small antenna, the microwaves are directed into the cooking compartment of the oven, and, as they enter, they pass a set of turning metal fan blades. This is the stirrer, which disperses the microwaves uniformly over the surface of the food to be heated. As a microwave strikes a water molecule, resonance causes the molecule to align with the direction of the wave. An oscillating magnetron, a tube that generates radio waves, causes the microwaves to oscillate as well, and this, in turn, compels the water molecules to do the same. Thus, the water molecules are shifting in position several million times a second, and this vibration generates energy that heats the water.

Microwave ovens do not heat food from the inside out: like a conventional oven, they can only cook from the outside in. But so much energy is transferred to the water molecules that conduction does the rest, ensuring relatively uniform heating of the food. Incidentally, the resonance between microwaves and water molecules explains why many materials used in cooking dishes—materials that do not contain water—can be placed in a microwave oven without being melted or burned. Yet metal, though it also contains no water, is unsafe.

Metals have free electrons, which makes them good electrical conductors, and the presence of these free electrons means that the microwaves produce electric currents in the surfaces of metal objects placed in the oven. Depending on the shape of the object, these currents can jump, or arc, between points on the surface, thus producing sparks. On the other hand, the interior of the microwave oven itself is in fact metal, and this is so precisely because microwaves do bounce back and forth off of metal. Because the walls are flat and painted, however, currents do not arc between them.

Resonance of Sound Waves

A highly trained singer can hit a note that causes a wine glass to shatter, but what causes this to happen is not the frequency of the note, per se. In other words, the shattering is not necessarily because of the fact that the note is extremely high; rather, it is due to the phenomenon of resonance. The natural, or resonant, frequency in the wine glass, as with all objects, is determined by its shape and composition. If the singer's voice (or a note from an instrument) hits the resonant frequency, there will be a transfer of energy, as with the father pushing his daughter on the swing. In this case, however, a full transfer of energy from the voice or musical instrument can overload the glass, causing it to shatter.

Another example of resonance and sound waves is feedback, popularized in the 1960s by rock guitarists such as Jimi Hendrix and Pete Townsend of the Who. When a musician strikes a note on an electric guitar string, the string oscillates, and an electromagnetic device in the guitar converts this oscillation into an electrical pulse that it sends to an amplifier. The amplifier passes this oscillation on to the speaker, but if the frequency of the speaker is the same as that of the vibrations in the guitar, the result is feedback.

Both in scientific terms and in the view of a music fan, feedback adds energy. The feedback from the speaker adds energy to the guitar body, which, in turn, increases the energy in the vibration of the guitar strings and, ultimately, the power of the electrical signal is passed on to the amp. The result is increasing volume, and the feedback thus creates a loop that continues to repeat until the volume drowns out all other notes.

KEY TERMS

AMPLITUDE: The maximum displacement of particles from their normal position during a single period of oscillation.

CYCLE: One full repetition of oscillation.

FREQUENCY: For a particle experiencing oscillation, frequency is the number of cycles that take place during one second. Frequency is measured in Hertz.

HARMONIC MOTION: The repeated movement of a particle about a position of equilibrium, or balance.

HERTZ: A unit for measuring frequency, named after nineteenth-century German physicist Heinrich Rudolf Hertz (1857-1894). Higher frequencies are expressed in terms of kilohertz (kHz; 10^3 or 1,000 cycles per second) or megahertz (MHz; 10^6 or 1 million cycles per second.)

KINETIC ENERGY: The energy that an object possesses due to its motion, as with a sled when sliding down a hill. This is contrasted with potential energy.

LONGITUDINAL WAVE: A wave in which the movement of vibration is in the same direction as the wave itself. This is contrasted to a transverse wave.

MAXIMUM DISPLACEMENT: For an object in oscillation, maximum displacement is the furthest point from stable equilibrium.

OSCILLATION: A type of harmonic motion, typically periodic, in one or more dimensions.

PERIOD: The amount of time required for one cycle in oscillating motion.

PERIODIC MOTION: Motion that is repeated at regular intervals. These intervals are known as periods.

PERIODIC WAVE: A wave in which a uniform series of crests and troughs follow one after the other in regular succession.

POTENTIAL ENERGY: The energy that an object possesses due to its position, as, for instance, with a sled at the top of a hill. This is contrasted with kinetic energy.

RESONANCE: The condition in which force is applied to an object in oscillation at the point of maximum amplitude.

RESONANT FREQUENCY: A frequency that matches that of an oscillating object.

STABLE EQUILIBRIUM: A position in which, if an object were disturbed, it would tend to return to its original position. For an object in oscillation, stable equilibrium is in the middle of a cycle, between two points of maximum displacement.

TRANSVERSE WAVE: A wave in which the vibration or motion is perpendicular to the direction in which the wave is moving. This is contrasted to a longitudinal wave.

WAVE MOTION: A type of harmonic motion that carries energy from one place to another without actually moving any matter.

How Resonance Can Break a Bridge

The power of resonance goes beyond shattering a glass or torturing eardrums with feedback; it can actually destroy large structures. There is an old folk saying that a cat can destroy a bridge if it walks across it in a certain way. This may or may not be true, but it is certainly conceivable that a group of soldiers marching across a bridge can cause it to crumble, even though it is capable of holding much more than their weight, if the rhythm of their synchronized footsteps resonates with the natural frequency of the bridge. For this reason, officers or sergeants typically order their troops to do something very unmilitary—to march out of step—when crossing a bridge.

The resonance between vibrations produced by wind and those of the structure itself brought down a powerful bridge in 1940, a highly dramatic illustration of physics in action that was captured on both still photographs and film. Located on Puget Sound near Seattle, Washington, the Tacoma Narrows Bridge was, at 2,800 ft (853 m) in length, the third-longest suspension bridge in the world. But on November 7, 1940, it gave way before winds of 42 mi (68 km) per hour.

It was not just the speed of these winds, but the fact that they produced oscillations of resonant frequency, that caused the bridge to twist and, ultimately, to crumble. In those few seconds of battle with the forces of nature, the bridge writhed and buckled until a large segment collapsed into the waters of Puget Sound. Fortunately, no one was killed, and a new, more stable bridge was later built in place of the one that had come to be known as "Galloping Gertie." The incident led to increased research and progress in understanding of aerodynamics, harmonic motion, and resonance.

Where to Learn More

Beiser, Arthur. *Physics,* 5th ed. Reading, MA: Addison-Wesley, 1991.

Berger, Melvin. *The Science of Music.* Illustrated by Yvonne Buchanan. New York: Crowell, 1989.

"*Bridges and Resonance*" (Web site). <http://instruction.ferris.edu/loub/media/BRIDGE/Bridge.html> (April 23, 2001).

"*Resonance*" (Web site). <http://hyperphysics.phyastr.gsu.edu/hbase/sound/reson.html> (April 26, 2001).

"*Resonance*" (Web site). <http://www.exploratorium.edu/xref/phenomena/resonance.html> (April 23, 2001).

"*Resonance.*" *The Physics Classroom* (Web site). <http://www.glenbrook.k12.il.us/gbssci/phys/Class/sound/u11l5a.html> (April 26, 2001).

"*Resonance Experiment*" (Web site). <http://131.123.17.138/> (April 26, 2001).

"*Resonance, Frequency, and Wavelength*" (Web site). <http://www.cpo.com/CPOCatalog/SW/sw_b1.html> (April 26, 2001).

Suplee, Curt. *Everyday Science Explained.* Washington, D.C.: National Geographic Society, 1996.

"*Tacoma Narrows Bridge Disaster*" (Web site). <http://www.enm.bris.ac.uk/research/nonlinear/tacoma/tacoma.html> (April 23, 2001).

INTERFERENCE

CONCEPT

When two or more waves interact and combine, they interfere with one another. But interference is not necessarily bad: waves may interfere constructively, resulting in a wave larger than the original waves. Or, they may interfere destructively, combining in such a way that they form a wave smaller than the original ones. Even so, destructive interference may have positive effects: without the application of destructive interference to the muffler on an automobile exhaust system, for instance, noise pollution from cars would be far worse than it is. Other examples of interference, both constructive and destructive, can be found wherever there are waves: in water, in sound, in light.

HOW IT WORKS

Waves

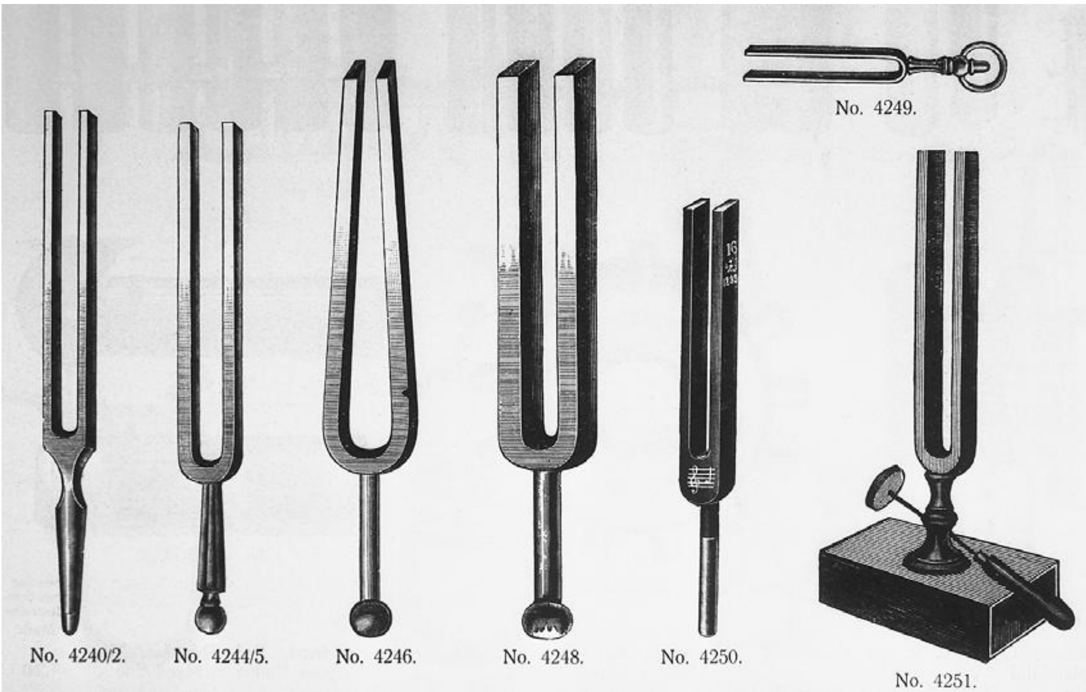
Whenever energy ripples through space, there is a wave. In fact, wave motion can be defined as a type of harmonic motion (repeated movement of a particle about a position of equilibrium, or balance) that carries energy from one place to another without actually moving any matter. A wave on the ocean is an example of a mechanical wave, or one that involves matter; but though the matter moves in place, it is only the energy in the wave that experiences net movement.

Wave motion is related to oscillation, a type of harmonic motion in one or more dimensions. There is one critical difference, however: oscillation involves no net movement, only movement in place, whereas the harmonic motion of waves carries energy from one place to another. Yet, individual waves themselves are oscillating even as the overall wave pattern moves.

A transverse wave forms a regular up-and-down pattern in which the oscillation is perpendicular to the direction the wave is moving. Ocean waves are transverse, though they also have properties of longitudinal waves. In a longitudinal wave, of which a sound wave is the best example, oscillation occurs in the same direction as the wave itself.

PARAMETERS OF WAVE MOTION. Some waves, composed of pulses, do not follow regular patterns. However, the waves of principal concern in the present context are periodic waves, ones in which a uniform series of crests and troughs follow each other in regular succession. Periodic motion is movement repeated at regular intervals called periods. In the case of wave motion, a period (represented by the symbol T) is the amount of time required to complete one full cycle of the wave, from trough to crest and back to trough.

Period can be mathematically related to several other aspects of wave motion, including wave speed, frequency, and wavelength. Frequency (abbreviated f) is the number of waves passing through a given point during the interval of one second. It is measured in Hertz (Hz), named after nineteenth-century German physicist Heinrich Rudolf Hertz (1857-1894), and a Hertz is equal to one cycle of oscillation per second. Higher frequencies are expressed in terms of kilohertz (kHz; 10^3 or 1,000 cycles per second) or megahertz (MHz; 10^6 or 1 million cycles per second.) Wavelength (represented by the symbol abbreviated λ, the Greek letter lambda) is the distance between a crest and the adjacent crest, or a



A PIANO TUNER, USING A TUNING FORK SUCH AS THE ONES SHOWN ABOVE, UTILIZES INTERFERENCE TO TUNE THE INSTRUMENT. *(Bettmann/Corbis. Reproduced by permission.)*

trough and an adjacent trough, of a wave. The higher the frequency, the shorter the wavelength.

Another parameter for describing wave motion—one that is mathematically independent from the quantities so far described—is amplitude, or the maximum displacement of particles from a position of stable equilibrium. For an ocean wave, amplitude is the distance from either the crest or the trough to the level that the ocean would maintain if it were perfectly still.

SUPERPOSITION AND INTERFERENCE

SUPERPOSITION. The principle of superposition holds that when several individual but similar physical events occur in close proximity, the resulting effect is the sum of the magnitude of the separate events. This is akin to the popular expression, "The whole is greater than the sum of the parts," and it has numerous applications in physics.

Where the strength of a gravitational field is being measured, for instance, superposition dictates that the strength of that field at any given point is the sum of the mass of the individual particles in that field. In the realm of electromagnetic force, the same statement applies, though the units being added are electrical charges or magnetic poles, rather than quantities of mass. Likewise, in an electrical circuit, the total current or voltage is the sum of the individual currents and voltages in that circuit.

Superposition applies only in equations for linear events—that is, phenomena that involve movement along a straight line. Waves are linear phenomena, and, thus, the principle describes the behavior of all waves when they come into contact with one another. If two or more waves enter the same region of space at the same time, then, at any instant, the total disturbance produced by the waves at any point is equal to the sum of the disturbances produced by the individual waves.

INTERFERENCE. The principle of superposition does not require that waves actually combine; rather, the net effect is as though they were combined. The actual combination or joining of two or more waves at a given point in space is called interference, and, as a result, the waves produce a single wave whose properties are determined by the properties of the individual waves.

If two waves of the same wavelength occupy the same space in such a way that their crests and

IF THIS BOAT'S WAKE WERE TO CROSS THE WAKE OF ANOTHER BOAT, THE RESULT WOULD BE BOTH CONSTRUCTIVE AND DESTRUCTIVE INTERFERENCE. *(Photograph by Roger Ressmeyer/Corbis. Reproduced by permission.)*

troughs align, the wave they produce will have an amplitude greater than that possessed by either wave initially. This is known as constructive interference. The more closely the waves are in phase—that is, perfectly aligned—the more constructive the interference.

It is also possible that two or more waves can come together such that the trough of one meets the crest of the other, or vice versa. In this case, what happens is destructive interference, and the resulting amplitude is the difference between the values for the individual waves. If the waves are perfectly unaligned—in other words, if the trough of one exactly meets the crest of the other—their amplitudes cancel out, and the result is no wave at all.

Resonance

It is easy to confuse interference with resonance, and, therefore, a word should be said about the latter phenomenon. The term resonance describes a situation in which force is applied to an oscillator at the point of maximum amplitude. In this way, the motion of the outside force is perfectly matched to that of the oscillator, making possible a transfer of energy. As with interference, resonance implies alignment between two physical entities; however, there are several important differences.

Resonance can involve waves, as, for instance, when sound waves resonate with the vibrations of an oscillator, causing a transfer of energy that sometimes produces dramatic results. (See essay on Resonance.) But in these cases, a wave is interacting with an oscillator, not a wave with a wave, as in situations of interference. Furthermore, whereas resonance entails a transfer of energy, interference involves a combination of energy.

TRANSFER VS. COMBINATION. The importance of this distinction is easy to see if one substitutes money for energy, and people for objects. If one passes on a sum of money to another person, a business, or an institution—as a loan, repayment of a loan, a purchase, or a gift—this is an example of a transfer. On the other hand, when married spouses each earn paychecks, their cash is combined.

Transfer thus indicates that the original holder of the cash (or energy) no longer has it. Yet, if the holder of the cash combines funds with those of another, both share rights to an amount of money greater than the amount each originally owned. This is analogous to constructive interference.

On the other hand, a husband and wife (or any other group of people who pool their cash) also share liabilities, and, thus, a married person may be subject to debt incurred by his or her spouse. If one spouse creates debt so great that the other spouse cannot earn enough to maintain the payments, this painful situation is analogous to destructive interference.

REAL-LIFE APPLICATIONS

Mechanical Waves

One of the easiest ways to observe interference is by watching the behavior of mechanical waves. Drop a stone into a still pond, and watch how its waves ripple: this, as with most waveforms in water, is an example of a surface wave, or one that displays aspects of both transverse and longitudinal wave motion. Thus, as the concentric circles of a longitudinal wave ripple outward in one dimension, there are also transverse movements along a plane perpendicular to that of the longitudinal wave.

While the first wave is still rippling across the water, drop another stone close to the place where the first one was dropped. Now, there are two surface waves, crests and troughs colliding and interfering. In some places, they will interfere constructively, producing a wave—or rather, a portion of a wave—that is greater in amplitude than either of the original waves. At other places, there will be destructive interference, with some waves so perfectly out of phase that at one instant in time, a given spot on the water may look as though it had not been disturbed at all.

One of the interesting aspects of this interaction is the lack of uniformity in the instances of interference. As suggested in the preceding paragraph, it is usually not entire waves, but merely portions of waves, that interfere constructively or destructively. The result is that a seemingly simple event—dropping two stones into a still pond—produces a dazzling array of colliding circles, broken by outwardly undisturbed areas of destructive interference.

A similar phenomenon, though manifested by the interaction of geometric lines rather than concentric circles, occurs when two power boats pass each other on a lake. The first boat chops up the water, creating a wake that widens behind it: when seen from the air, the boat appears to be at the apex of a triangle whose sides are formed by rippling eddies of water.

Now, another boat passes through the wake of the first, only it is going in the opposite direction and producing its own ever-widening wake as it goes. As the waves from the two boats meet, some are in phase, but, more often than not, they are only partly in phase, or they possess differing wavelengths. Therefore, the waves at least partially cancel out one another in places, and in other places, reinforce one another. The result is an interesting patchwork of patterns when seen from the air.

Sound Waves

IN TUNE AND OUT OF TUNE. The relationships between musical notes can be intriguing, and though tastes in music vary, most people know when music is harmonious and when it is discordant. As discussed in the essay on frequency, this harmony or discord can be equated to the mathematical relationships between the frequencies of specific notes: the lower the numbers involved in the ratio, the more pleasing the sound.

The ratio between the frequency of middle C and that of its first harmonic—that is, the C note exactly one octave above it—is a nice, clean 1:2. If one were to play a song in the key of C-which, on a piano, involves only the "white notes" C-D-E-F-G-A-B—everything should be perfectly harmonious and (presumably) pleasant to the ear. But what if the piano itself is out of tune? Or what if one key is out of tune with the others?

The result, for anyone who is not tone-deaf, produces an overall impression of unpleasantness: it might be a bit hard to identify the source of this discomfort, but it is clear that something is amiss. At best, an out-of-tune piano might sound like something that belonged in a saloon from an old Western; at worst, the sound of notes that do not match their accustomed frequencies can be positively grating.

HOW A TUNING FORK WORKS. To rectify the situation, a professional piano tuner uses a tuning fork, an instrument that produces a single frequency—say, 264 Hz, which is the frequency of middle C. The piano tuner strikes the tuning fork, and at the same time strikes the appropriate key on the piano. If their frequencies are perfectly aligned, so is the sound of both; but, more likely, there will be interference, both constructive and destructive.

As time passes—measured in seconds or even fractions of seconds—the sounds of the tuning fork and that of the piano key will alternate between constructive and destructive interference. In the case of constructive interference, their combined sound will become louder than the individual sounds of either; and when the interference is destructive, the sound of both together will be softer than that produced by either the fork or the key.

The piano tuner listens for these fluctuations of loudness, which are called beats, and adjusts the tension in the appropriate piano string until the beats disappear completely. As long as there are beats, the piano string and the tuning fork will produce together a frequency that is the average of the two: if, for instance, the out-of-tune middle C string vibrates at 262 Hz, the resulting frequency will be 263 Hz.

DIFFERENCE TONES. Another interesting aspect of the interaction between notes is the "difference tone," created by discord, which the human ear perceives as a third tone. Though E and F are both part of the C scale, when struck together, the sound is highly discordant. In light of what was said above about ratios between frequencies, this dissonance is fitting, as the ratio here involves relatively high numbers—15:16.

When two notes are struck together, they produce a combination tone, perceived by the human ear as a third tone. If the two notes are harmonious, the "third tone" is known as a summation tone, and is equal to the combined frequencies of the two notes. But if the combination is dissonant, as in the case of E and F, the third tone is known as a difference tone, equal to the difference in frequencies. Since an E note vibrates at 330 Hz, and an F note at 352 Hz, the resulting difference tone is equal to 22 Hz.

DESTRUCTIVE INTERFERENCE IN SOUND WAVES. When music is played in a concert hall, it reverberates off the walls of the auditorium. Assuming the place is well designed acoustically, these bouncing sound waves will interfere constructively, and the auditorium comes alive with the sound of the music. In other situations, however, the sound waves may interfere destructively, and the result is a certain muffled deadness to the sound.

Clearly, in a music hall, destructive interference is a problem; but there are cases in which it can be a benefit—situations, that is, in which the purpose, indeed, is to deaden the sound. One example is an automobile muffler. A car's exhaust system makes a great deal of noise, and, thus, if a car does not have a proper muffler, it creates a great deal of noise pollution. A muffler counteracts this by producing a sound wave out of phase with that of the exhaust system; hence, it cancels out most of the noise.

Destructive interference can also be used to reduce sound in a room. Once again, a machine is calibrated to generate sound waves that are perfectly out of phase with the offending noise—say, the hum of another machine. The resulting effect conveys the impression that there is no noise in the room, though, in fact, the sound waves are still there; they have merely canceled each other out.

ELECTROMAGNETIC WAVES

In 1801, English physicist Thomas Young (1773-1829), known for Young's modulus of elasticity became the first scientist to identify interference in light waves. Challenging the corpuscular theory of light put forward by Sir Isaac Newton (1642-1727), Young set up an experiment in which a beam of light passed through two closely spaced pinholes onto a screen. If light was truly made of particles, he said, the beams would project two distinct points onto the screen. Instead, what he saw was a pattern of interference.

In fact, Newton was partly right, but Young's discovery helped advance the view of light as a wave, which is also partly right. (According to quantum theory, developed in the twentieth century, light behaves both as waves and as particles.) The interference in the visible spectrum that Young witnessed was manifested as bright and dark bands. These bands are known as fringes—variations in intensity not unlike the beats created in some instances of sound interference, described above.

OILY FILMS AND RAINBOWS. Many people have noticed the strangely beautiful pattern of colors generated when light interacts with an oily substance, as when light reflected on a soap bubble produces an astonishing array of shades. Sometimes, this can happen in situations not otherwise aesthetically pleasing: an oily film in a parking lot, left there by a car's leaky

KEY TERMS

AMPLITUDE: The maximum displacement of particles in oscillation from a position of stable equilibrium. For an ocean wave, amplitude is the distance from either the crest or the trough to the level that the ocean would maintain if it were perfectly still.

CONSTRUCTIVE INTERFERENCE: A type of interference that occurs when two or more waves combine in such a way that they produce a wave whose amplitude is greater than that of the original waves. If waves are perfectly in phase—in other words, if the crest and trough of one exactly meets the crest and trough of the other—then the resulting amplitude is the sum of the individual amplitudes of the separate waves.

CYCLE: In oscillation, a cycle occurs when the oscillating particle moves from a certain point in a certain direction, then switches direction and moves back to the original point. Typically, this is from the position of stable equilibrium to maximum displacement and back again to the stable equilibrium position. In a wave, a cycle is equivalent to the movement from trough to crest and back to trough.

DESTRUCTIVE INTERFERENCE: A type of interference that occurs when two or more waves combine to produce a wave whose amplitude is less than that of the original waves. If waves are perfectly out of phase—in other words, if the trough of one exactly meets the crest of the other, and vice versa—their amplitudes cancel out, and the result is no wave at all.

FREQUENCY: In wave motion, frequency is the number of waves passing through a given point during the interval of one second. The higher the frequency, the shorter the wavelength. Frequency is mathematically related to wave speed and period.

HARMONIC MOTION: The repeated movement of a particle about a position of equilibrium, or balance.

HERTZ: A unit for measuring frequency, named after nineteenth-century German physicist Heinrich Rudolf Hertz (1857-1894). High frequencies are expressed in terms of kilohertz (kHz; 10^3 or 1,000 cycles per second) or megahertz (MHz; 10^6 or 1 million cycles per second.)

INTERFERENCE: The combination of two or more waves at a given point in space to produce a wave whose properties are determined by the properties of the individual waves. This accords with the principle of superposition.

LONGITUDINAL WAVE: A wave in which the movement of vibration is in the same direction as the wave itself. This is contrasted to a transverse wave.

MAXIMUM DISPLACEMENT: For an object in oscillation, maximum displacement is the furthest point from stable equilibrium.

MECHANICAL WAVE: A type of wave—for example, a wave on the ocean—that involves matter. The matter itself may move in place, but as with all types of wave motion, there is no net movement of matter—only of energy.

OSCILLATION: A type of harmonic motion, typically periodic, in one or more dimensions.

KEY TERMS CONTINUED

PERIOD: For wave motion, a period is the amount of time required to complete one full cycle. Period is mathematically related to frequency, wavelength, and wave speed.

PERIODIC MOTION: Motion that is repeated at regular intervals. These intervals are known as periods.

PERIODIC WAVE: A wave in which a uniform series of crests and troughs follow one after the other in regular succession.

PHASE: When two waves of the same frequency and amplitude are perfectly aligned, they are said to be in phase.

PRINCIPLE OF SUPERPOSITION: A physical principle stating that when several individual, but similar, physical events occur in close proximity to one another, the resulting effect is the sum of the magnitude of the separate events. Interference is an example of superposition.

PULSE: An isolated, non-periodic disturbance that takes place in wave motion of a type other than that of a periodic wave.

RESONANCE: The condition in which force is applied to an object in oscillation at the point of maximum amplitude.

STABLE EQUILIBRIUM: A position in which, if an object were disturbed, it would tend to return to its original position. For an object in oscillation, stable equilibrium is in the middle of a cycle, between two points of maximum displacement.

SURFACE WAVE: A wave that exhibits the behavior of both a transverse wave and a longitudinal wave.

TRANSVERSE WAVE: A wave in which the vibration or motion is perpendicular to the direction in which the wave is moving. This is contrasted to a longitudinal wave.

WAVELENGTH: The distance between a crest and the adjacent crest, or the trough and an adjacent trough, of a wave. Wavelength, abbreviated λ (the Greek letter lambda) is mathematically related to wave speed, period, and frequency.

WAVE MOTION: A type of harmonic motion that carries energy from one place to another, without actually moving any matter.

crankcase, can produce a rainbow of colors if the sunlight hits it just right.

This happens because the thickness of the oil causes a delay in reflection of the light beam. Some colors pass through the film, becoming delayed and, thus, getting out of phase with the reflected light on the surface of the film. These shades destructively interfere to such an extent that the waves are cancelled, rendering them invisible. Other colors reflect off the surface so that they are perfectly in phase with the light traveling through the film, and appear as an attractive swirl of color on the surface of the oil.

The phenomenon of light-wave interference with oily or filmy surfaces has the effect of filtering light, and, thus, has a number of applications in areas relating to optics: sunglasses, lenses for binoculars or cameras, and even visors for astronauts. In each case, unfiltered light could be harmful or, at least, inconvenient for the user, and the destructive interference eliminates certain colors and unwanted reflections.

RADIO WAVES. Visible light is only a small part of the electromagnetic spectrum, whose broad range of wave phenomena are, likewise, subject to constructive or destructive interference. After visible light, the area of the spectrum most people experience during an average day is the realm of relatively low-frequency, long-wavelength radio waves and microwaves, the latter including television broadcast signals.

People who rely on an antenna for their TV reception are likely to experience interference at some point. However, an increasing number of Americans use either cable or satellite systems to pick up TV programs. These are much less susceptible to interference, due to the technology of coaxial cable, on the one hand, and digital compression, on the other. Thus, interference in television reception is a gradually diminishing problem.

Interference among radio signals continues to be a challenge, since most people still hear the radio via old-fashioned means rather than through new technology, such as the Internet. A number of interference problems are created by activity on the Sun, which has an enormously powerful electromagnetic field. Obviously, such interference is beyond the control of most radio listeners, but according to a Web page set up by WHKY Radio in Hickory, North Carolina, there are a number of things listeners can do to decrease interference in their own households.

Among the suggestions offered at the WHKY Web site is this: "Nine times out of ten, if your radio is near a computer, it will interfere with your radio. Computers send out all kinds of signals that your radio 'thinks' is a real radio signal. Try to locate your radio away from computers... especially the monitor." The Web site listed a number of other household appliances, as well as outside phenomena such as power lines or thunderstorms, that can contribute to radio interference.

WHERE TO LEARN MORE

Beiser, Arthur. Physics, 5th ed. Reading, MA: Addison-Wesley, 1991.

Bloomfield, Louis A. *"How Things Work: Radio." How Things Work* (Web site). <http://rabi.phys.virginia.edu/HTW//radio.html> (April 27, 2001).

Harrison, David. *"Sound"* (Web site). <http://www.newi.ac.uk/buckley/sound.html> (April 27, 2001).

Interference Handbook/Federal Communications Commission (Web site). <http://www.fcc.gov/cib/Publications/tvibook.html> (April 27, 2001).

Internet Resources for Sound and Light (Web site). <http://electro.sau.edu/SLResources.html> (April 25, 2001).

"Light—A-to-Z Science." DiscoverySchool.com (Web site). <http://school.discovery.com/homeworkhelp/worldbook/atozscience/l/323260.html> (April 27, 2001).

Oxlade, Chris. *Light and Sound.* Des Plaines, IL: Heinemann Library, 2000.

"Sound Wave—Constructive and Destructive Interference" (Web site). <http://csgrad.cs.vt.edu/~chin/interference.html> (April 27, 2001).

Topp, Patricia. *This Strange Quantum World and You.* Nevada City, CA: Blue Dolphin, 1999.

WHKY Radio and TV (Web site). <http://www.whky.com/antenna.html> (April 27, 2001).

DIFFRACTION

CONCEPT

Diffraction is the bending of waves around obstacles, or the spreading of waves by passing them through an aperture, or opening. Any type of energy that travels in a wave is capable of diffraction, and the diffraction of sound and light waves produces a number of effects. (Because sound waves are much larger than light waves, however, diffraction of sound is a part of daily life that most people take for granted.) Diffraction of light waves, on the other hand, is much more complicated, and has a number of applications in science and technology, including the use of diffraction gratings in the production of holograms.

HOW IT WORKS

Comparing Sound and Light Diffraction

Imagine going to a concert hall to hear a band, and to your chagrin, you discover that your seat is directly behind a wide post. You cannot see the band, of course, because the light waves from the stage are blocked. But you have little trouble hearing the music, since sound waves simply diffract around the pillar. Light waves diffract slightly in such a situation, but not enough to make a difference with regard to your enjoyment of the concert: if you looked closely while sitting behind the post, you would be able to observe the diffraction of the light waves glowing slightly, as they widened around the post.

Suppose, now, that you had failed to obtain a ticket, but a friend who worked at the concert venue arranged to let you stand outside an open door and hear the band. The sound quality would be far from perfect, of course, but you would still be able to hear the music well enough. And if you stood right in front of the doorway, you would be able to see light from inside the concert hall. But, if you moved away from the door and stood with your back to the building, you would see little light, whereas the sound would still be easily audible.

WAVELENGTH AND DIFFRACTION. The reason for the difference—that is, why sound diffraction is more pronounced than light diffraction—is that sound waves are much, much larger than light waves. Sound travels by longitudinal waves, or waves in which the movement of vibration is in the same direction as the wave itself. Longitudinal waves radiate outward in concentric circles, rather like the rings of a bull's-eye.

The waves by which sound is transmitted are larger, or comparable in size to, the column or the door—which is an example of an aperture—and, hence, they pass easily through apertures and around obstacles. Light waves, on the other hand, have a wavelength, typically measured in nanometers (nm), which are equal to one-millionth of a millimeter. Wavelengths for visible light range from 400 (violet) to 700 nm (red): hence, it would be possible to fit about 5,000 of even the longest visible-light wavelengths on the head of a pin!

Whereas differing wavelengths in light are manifested as differing colors, a change in sound wavelength indicates a change in pitch. The higher the pitch, the greater the frequency, and, hence, the shorter the wavelength. As with light waves—though, of course, to a much lesser

Holograms are made possible through the principle of diffraction. Shown here is a hologram of a space shuttle orbiting Earth. *(Photograph by Roger Ressmeyer/Corbis. Reproduced by permission.)*

extent—short-wavelength sound waves are less capable of diffracting around large objects than are long-wavelength sound waves. Chances are, then, that the most easily audible sounds from inside the concert hall are the bass and drums; higher-pitched notes from a guitar or other instruments, such as a Hammond organ, are not as likely to reach a listener outside.

Observing Diffraction in Light

Due to the much wider range of areas in which light diffraction has been applied by scientists, diffraction of light and not sound will be the principal topic for the remainder of this essay. We have already seen that wavelength plays a role in diffraction; so, too, does the size of the aperture relative to the wavelength. Hence, most studies of diffraction in light involve very small openings, as, for instance, in the diffraction grating discussed below.

But light does not only diffract when passing through an aperture, such as the concert-hall door in the earlier illustration; it also diffracts around obstacles, as, for instance, the post or pillar mentioned earlier. This can be observed by looking closely at the shadow of a flagpole on a bright morning. At first, it appears that the shadow is "solid," but if one looks closely enough, it becomes clear that, at the edges, there is a blur-

The checkout scanners in grocery stores use holographic technology that can read a universal product code (UPC) from any angle. *(Photograph by Bob Rowan; Progressive Image/Corbis. Reproduced by permission.)*

ring from darkness to light. This "gray area" is an example of light diffraction.

Where the aperture or obstruction is large compared to the wave passing through or around it, there is only a little "fuzziness" at the edge, as in the case of the flagpole. When light passes through an aperture, most of the beam goes straight through without disturbance, with only the edges experiencing diffraction. If, however, the size of the aperture is close to that of the wavelength, the diffraction pattern will widen. Sound waves diffract at large angles through an open door, which, as noted, is comparable in size to a sound wave; similarly, when light is passed through extremely narrow openings, its diffraction is more noticeable.

Early Studies in Diffraction

Though his greatest contributions lay in his epochal studies of gravitation and motion, Sir Isaac Newton (1642-1727) also studied the production and propagation of light. Using a prism, he separated the colors of the visible light spectrum—something that had already been done by other scientists—but it was Newton who discerned that the colors of the spectrum could be recombined to form white light again.

Newton also became embroiled in a debate as the nature of light itself—a debate in which diffraction studies played an important role. Newton's view, known at the time as the corpuscular theory of light, was that light travels as a stream of particles. Yet, his contemporary, Dutch physicist and astronomer Christiaan Huygens (1629-1695), advanced the wave theory, or the idea that light travels by means of waves. Huygens maintained that a number of factors, including the phenomena of reflection and refraction, indicate that light is a wave. Newton, on the other hand, challenged wave theorists by stating that if light were actually a wave, it should be able to bend around corners—in other words, to diffract.

GRIMALDI IDENTIFIES DIFFRACTION. Though it did not become widely known until some time later, in 1648—more than a decade before the particle-wave controversy erupted—Johannes Marcus von Kronland (1595-1667), a scientist in Bohemia (now part of the Czech Republic), discovered the diffraction of light waves. However, his findings were not recognized until some time later; nor did he give a name to the phenomenon he had observed. Then, in 1660, Italian physicist Francesco Grimaldi (1618-1663) conducted an experiment with diffraction that gained widespread attention.

Grimaldi allowed a beam of light to pass through two narrow apertures, one behind the other, and then onto a blank surface. When he did so, he observed that the band of light hitting the surface was slightly wider than it should be, based on the width of the ray that entered the first aperture. He concluded that the beam had been bent slightly outward, and gave this phenomenon the name by which it is known today: diffraction.

FRESNEL AND FRAUNHOFER DIFFRACTION. Particle theory continued to have its adherents in England, Newton's homeland, but by the time of French physicist Augustin Jean Fresnel (1788-1827), an increasing number of scientists on the European continent had come to accept the wave theory. Fresnel's work, which he published in 1818, served to advance that theory, and, in particular, the idea of light as a transverse wave.

In *Memoire sur la diffraction de la lumiere,* Fresnel showed that the transverse-wave model accounted for a number of phenomena, including diffraction, reflection, refraction, interference, and polarization, or a change in the oscillation patterns of a light wave. Four years after publishing this important work, Fresnel put his ideas into action, using the transverse model to create a pencil-beam of light that was ideal for lighthouses. This prism system, whereby all the light emitted from a source is refracted into a horizontal beam, replaced the older method of mirrors used since ancient times. Thus Fresnel's work revolutionized the effectiveness of lighthouses, and helped save lives of countless sailors at sea.

The term "Fresnel diffraction" refers to a situation in which the light source or the screen are close to the aperture; but there are situations in which source, aperture, and screen (or at least two of the three) are widely separated. This is known as Fraunhofer diffraction, after German physicist Joseph von Fraunhofer (1787-1826), who in 1814 discovered the lines of the solar spectrum (source) while using a prism (aperture). His work had an enormous impact in the area of spectroscopy, or studies of the interaction between electromagnetic radiation and matter.

REAL-LIFE APPLICATIONS

Diffraction Studies Come of Age

Eventually the work of Scottish physicist James Clerk Maxwell (1831-1879), German physicist Heinrich Rudolf Hertz (1857-1894), and others confirmed that light did indeed travel in waves. Later, however, Albert Einstein (1879-1955) showed that light behaves both as a wave and, in certain circumstances, as a particle.

In 1912, a few years after Einstein published his findings, German physicist Max Theodor Felix von Laue (1879-1960) created a diffraction grating, discussed below. Using crystals in his grating, he proved that x rays are part of the electromagnetic spectrum. Laue's work, which earned him the Nobel Prize in physics in 1914, also made it possible to measure the length of x rays, and, ultimately, provided a means for studying the atomic structure of crystals and polymers.

SCIENTIFIC BREAKTHROUGHS MADE POSSIBLE BY DIFFRACTION STUDIES. Studies in diffraction advanced during the early twentieth century. In 1926, English physicist J. D. Bernal (1901-1971) developed the Bernal chart, enabling scientists to deduce the crystal structure of a solid by analyzing photographs of x-ray diffraction patterns. A decade later, Dutch-American physical chemist Peter Joseph William Debye (1884-1966) won the Nobel Prize in Chemistry for his studies in the diffraction of x rays and electrons in gases, which advanced understanding of molecular structure. In 1937, a year after Debye's Nobel, two other scientists—American physicist Clinton Joseph Davisson (1881-1958) and English physicist George Paget Thomson (1892-1975)—won the Prize in Physics for their discovery that crystals can bring about the diffraction of electrons.

Also, in 1937, English physicist William Thomas Astbury (1898-1961) used x-ray diffraction to discover the first information concerning nucleic acid, which led to advances in the study of DNA (deoxyribonucleic acid), the building-blocks of human genetics. In 1952, English biophysicist Maurice Hugh Frederick Wilkins (1916-) and molecular biologist Rosalind Elsie Franklin (1920-1958) used x-ray diffraction to photograph DNA. Their work directly influenced a breakthrough event that followed a year later: the discovery of the double-helix or double-spiral model of DNA by American molecular biologists James D. Watson (1928-) and Francis Crick (1916-). Today, studies in DNA are at the frontiers of research in biology and related fields.

Diffraction Grating

Much of the work described in the preceding paragraphs made use of a diffraction grating, first developed in the 1870s by American physicist Henry Augustus Rowland (1848-1901). A diffraction grating is an optical device that consists of not one but many thousands of apertures: Rowland's machine used a fine diamond point to rule glass gratings, with about 15,000 lines per in (2.2 cm). Diffraction gratings today can have as many as 100,000 apertures per inch. The apertures in a diffraction grating are not mere holes, but extremely narrow parallel slits that transform a beam of light into a spectrum.

Each of these openings diffracts the light beam, but because they are evenly spaced and the same in width, the diffracted waves experience constructive interference. (The latter phenomenon, which describes a situation in which two or more waves combine to produce a wave of greater magnitude than either, is discussed in the essay on Interference.) This constructive interference pattern makes it possible to view components of the spectrum separately, thus enabling a scientist to observe characteristics ranging from the structure of atoms and molecules to the chemical composition of stars.

X-RAY DIFFRACTION. Because they are much higher in frequency and energy levels, x rays are even shorter in wavelength than visible light waves. Hence, for x-ray diffraction, it is necessary to have gratings in which lines are separated by infinitesimal distances. These distances are typically measured in units called an angstrom, of which there are 10 million to a millimeter. Angstroms are used in measuring atoms, and, indeed, the spaces between lines in an x-ray diffraction grating are comparable to the size of atoms.

When x rays irradiate a crystal—in other words, when the crystal absorbs radiation in the form of x rays—atoms in the crystal diffract the rays. One of the characteristics of a crystal is that its atoms are equally spaced, and, because of this, it is possible to discover the location and distance between atoms by studying x-ray diffraction patterns. Bragg's law—named after the father-and-son team of English physicists William Henry Bragg (1862-1942) and William Lawrence Bragg (1890-1971)—describes x-ray diffraction patterns in crystals.

Though much about x-ray diffraction and crystallography seems rather abstract, its application in areas such as DNA research indicates that it has numerous applications for improving human life. The elder Bragg expressed this fact in 1915, the year he and his son received the Nobel Prize in physics, saying that "We are now able to look ten thousand times deeper into the structure of the matter that makes up our universe than when we had to depend on the microscope alone." Today, physicists applying x-ray diffraction use an instrument called a diffractometer, which helps them compare diffraction patterns with those of known crystals, as a means of determining the structure of new materials.

Holograms

A hologram—a word derived from the Greek *holos,* "whole," and *gram,* "message"—is a three-dimensional (3-D) impression of an object, and the method of producing these images is known as holography. Holograms make use of laser beams that mix at an angle, producing an interference pattern of alternating bright and dark lines. The surface of the hologram itself is a sort of diffraction grating, with alternating strips of clear and opaque material. By mixing a laser beam and the unfocused diffraction pattern of an object, an image can be recorded. An illuminating laser beam is diffracted at specific angles, in accordance with Bragg's law, on the surfaces of the hologram, making it possible for an observer to see a three-dimensional image.

Holograms are not to be confused with ordinary three-dimensional images that use only visible light. The latter are produced by a method known as stereoscopy, which creates a single image from two, superimposing the images to create the impression of a picture with depth. Though stereoscopic images make it seem as though one can "step into" the picture, a hologram actually enables the viewer to glimpse the image from any angle. Thus, stereoscopic images can be compared to looking through the plate-glass window of a store display, whereas holo-

KEY TERMS

APERTURE: An opening.

DIFFRACTION: The bending of waves around obstacles, or the spreading of waves by passing them through an aperture.

ELECTROMAGNETIC SPECTRUM: The complete range of electromagnetic waves on a continuous distribution from a very low range of frequencies and energy levels, with a correspondingly long wavelength, to a very high range of frequencies and energy levels, with a correspondingly short wavelength. Included on the electromagnetic spectrum are long-wave and short-wave radio; microwaves; infrared, visible, and ultraviolet light; x rays, and gamma rays.

FREQUENCY: The number of waves passing through a given point during the interval of one second. The higher the frequency, the shorter the wavelength.

LONGITUDINAL WAVE: A wave in which the movement of vibration is in the same direction as the wave itself. A sound wave is an example of a longitudinal wave.

PRISM: A three-dimensional glass shape used for the diffusion of light rays.

PROPAGATION: The act or state of traveling from one place to another.

RADIATION: In a general sense, radiation can refer to anything that travels in a stream, whether that stream be composed of subatomic particles or electromagnetic waves.

REFLECTION: A phenomenon whereby a light ray is returned toward its source rather than being absorbed at the interface.

REFRACTION: The bending of a light ray that occurs when it passes through a dense medium, such as water or glass.

SPECTRUM: The continuous distribution of properties in an ordered arrangement across an unbroken range. Examples of spectra (the plural of "spectrum") include the colors of visible light, or the electromagnetic spectrum of which visible light is a part.

TRANSVERSE WAVE: A wave in which the vibration or motion is perpendicular to the direction in which the wave is moving.

WAVELENGTH: The distance between a crest and the adjacent crest, or the trough and an adjacent trough, of a wave. The shorter the wavelength, the higher the frequency.

grams convey the sensation that one has actually stepped into the store window itself.

DEVELOPMENTS IN HOLOGRAPHY. While attempting to improve the resolution of electron microscopes in 1947, Hungarian-English physicist and engineer Dennis Gabor (1900-1979) developed the concept of holography and coined the term "hologram." His work in this area could not progress by a great measure, however, until the creation of the laser in 1960. By the early 1960s, scientists were using lasers to create 3-D images, and in 1971, Gabor received the Nobel Prize in physics for the discovery he had made a generation before.

Today, holograms are used on credit cards or other identification cards as a security measure, providing an image that can be read by an optical scanner. Supermarket checkout scanners use holographic optical elements (HOEs), which can read a universal product code (UPC) from any angle. Use of holograms in daily life and scientific research is likely to increase as scientists find new applications: for instance, holographic images will aid the design of everything from bridges to automobiles.

HOLOGRAPHIC MEMORY. One of the most fascinating areas of research in the field of holography is holographic memory. Computers use a binary code, a pattern of ones and zeroes that is translated into an electronic pulse, but holographic memory would greatly extend the capabilities of computer memory systems. Unlike most images, a hologram is not simply the sum of its constituent parts: the data in a holographic image is contained in every part of the image, meaning that part of the image can be destroyed without a loss of data.

To bring the story full-circle, holographic memory calls to mind an idea advanced by a scientist who, along with Huygens, was one of Newton's great professional rivals, German mathematician and philosopher Gottfried Wilhelm Leibniz (1646-1716). Though Newton is usually credited as the father of calculus, Leibniz developed his own version of calculus at around the same time.

As a philosopher, Leibniz had apparently had a number of strange ideas, which made him the butt of jokes among some sectors of European intellectual society: hence, the French writer and thinker Voltaire (François-Marie Arouet; 1694-1778) satirized him with the character Dr. Pangloss in *Candide* (1759). Few of Leibniz's ideas were more bizarre than that of the monad: an elementary particle of existence that reflected the whole of the universe.

In advancing the concept of a monad, Leibniz was not making a statement after the manner of a scientist: there was no proof that monads existed, nor was it possible to prove this in any scientific way. Yet, a hologram appears to be very much like a manifestation of Leibniz's imagined monads, and both the hologram and the monad relate to a more fundamental aspect of life: human memory. Neurological research in the late twentieth century suggested that the structure of memory in the human mind is holographic. Thus, for instance, a patient suffering an injury affecting 90% of the brain experiences only a 10% memory loss.

WHERE TO LEARN MORE

Barrett, Norman S. *Lasers and Holograms.* New York: F. Watts, 1985.

"Bragg's Law and Diffraction: How Waves Reveal the Atomic Structure of Crystals" (Web site). <http://www.journey.sunysb.edu/ProjectJava/Bragg/home.html> (May 6, 2001).

Burkig, Valerie. *Photonics: The New Science of Light.* Hillside, N.J.: Enslow Publishers, 1986.

"Diffraction of Sound" (Web site). <http://hyperphysics.phy-astr.gsu.edu/hbase/sound/diffrac.html> (May 6, 2001).

Gardner, Robert. *Experimenting with Light.* New York: F. Watts, 1991.

Graham, Ian. *Lasers and Holograms.* New York: Shooting Star Press, 1993.

Holoworld: Holography, Lasers, and Holograms (Web site). <http://www.holoworld.com> (May 6, 2001).

Proffen, T. H. and R. B. Neder. *Interactive Tutorial About Diffraction* (Web site). <http://www.uniwuerzburg.de/mineralogie/crystal/teaching/teaching.html> (May 6, 2001).

Snedden, Robert. *Light and Sound.* Des Plaines, IL: Heinemann Library, 1999.

"Wave-Like Behaviors of Light." The Physics Classroom (Web site). <http://www.glenbrook.k12.il.us/gbssci/phys/Class/light/u12l1a.html> (May 6, 2001).

DOPPLER EFFECT

CONCEPT

Almost everyone has experienced the Doppler effect, though perhaps without knowing what causes it. For example, if one is standing on a street corner and an ambulance approaches with its siren blaring, the sound of the siren steadily gains in pitch as it comes closer. Then, as it passes, the pitch suddenly lowers perceptibly. This is an example of the Doppler effect: the change in the observed frequency of a wave when the source of the wave is moving with respect to the observer. The Doppler effect, which occurs both in sound and electromagnetic waves—including light waves—has a number of applications. Astronomers use it, for instance, to gauge the movement of stars relative to Earth. Closer to home, principles relating to the Doppler effect find application in radar technology. Doppler radar provides information concerning weather patterns, but some people experience it in a less pleasant way: when a police officer uses it to measure their driving speed before writing a ticket.

HOW IT WORKS

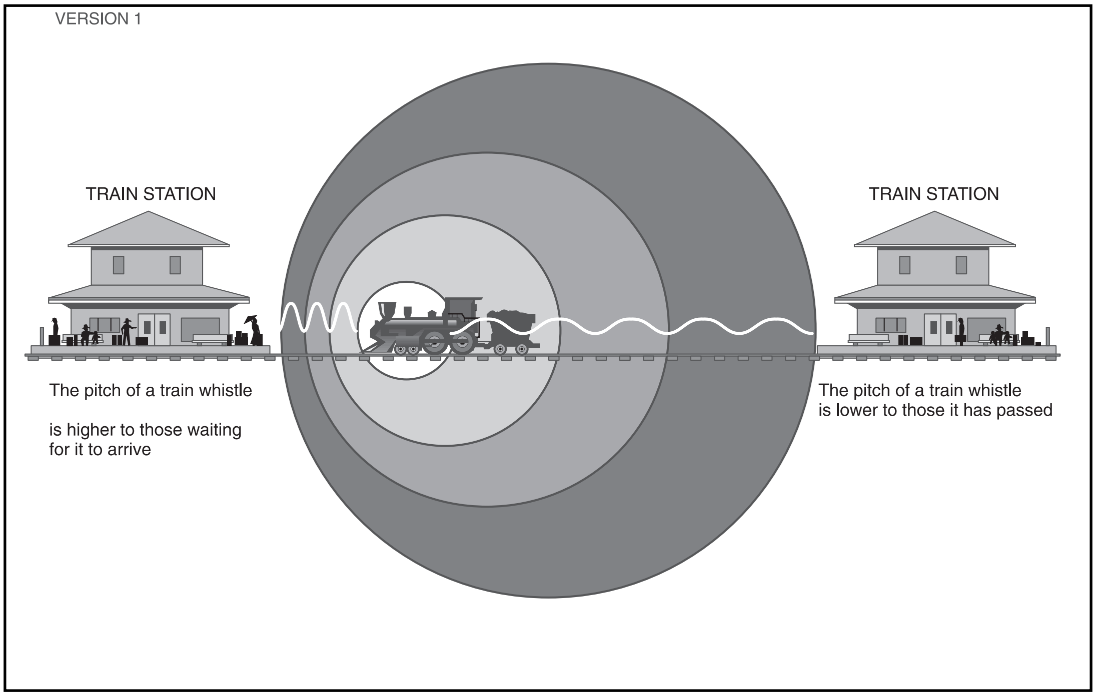
Wave Motion and Its Properties

Sound and light are both examples of energy, and both are carried on waves. Wave motion is a type of harmonic motion that carries energy from one place to another without actually moving any matter. It is related to oscillation, a type of harmonic motion in one or more dimensions. Oscillation involves no net movement, only movement in place; yet individual points in the wave medium are oscillating even as the overall wave pattern moves.

The term periodic motion, or movement repeated at regular intervals called periods, describes the behavior of periodic waves—waves in which a uniform series of crests and troughs follow each other in regular succession. A period (represented by the symbol T) is the amount of time required to complete one full cycle of the wave, from trough to crest and back to trough.

Period is mathematically related to several other aspects of wave motion, including wave speed, frequency, and wavelength. Frequency (abbreviated f) is the number of waves passing through a given point during the interval of one second. It is measured in Hertz (Hz), named after nineteenth-century German physicist Heinrich Rudolf Hertz (1857-1894), and a Hertz is equal to one cycle of oscillation per second. Higher frequencies are expressed in terms of kilohertz (kHz; 10^3 or 1,000 cycles per second); megahertz (MHz; 10^6 or 1 million cycles per second); and gigahertz (GHz; 10^9 or 1 billion cycles per second.) Wavelength (represented by the symbol λ, the Greek letter lambda) is the distance between a crest and the adjacent crest, or a trough and an adjacent trough, of a wave. The higher the frequency, the shorter the wavelength.

Amplitude, though mathematically independent from the parameters discussed, is critical to the understanding of sound. Defined as the maximum displacement of a vibrating material, amplitude is the "size" of a wave. The greater the amplitude, the greater the energy the wave contains: amplitude indicates intensity, which, in the case of sound waves, is manifested as what people commonly call "volume." Similarly, the



THE DOPPLER EFFECT.

amplitude of a light wave determines the intensity of the light.

FRAME OF REFERENCE

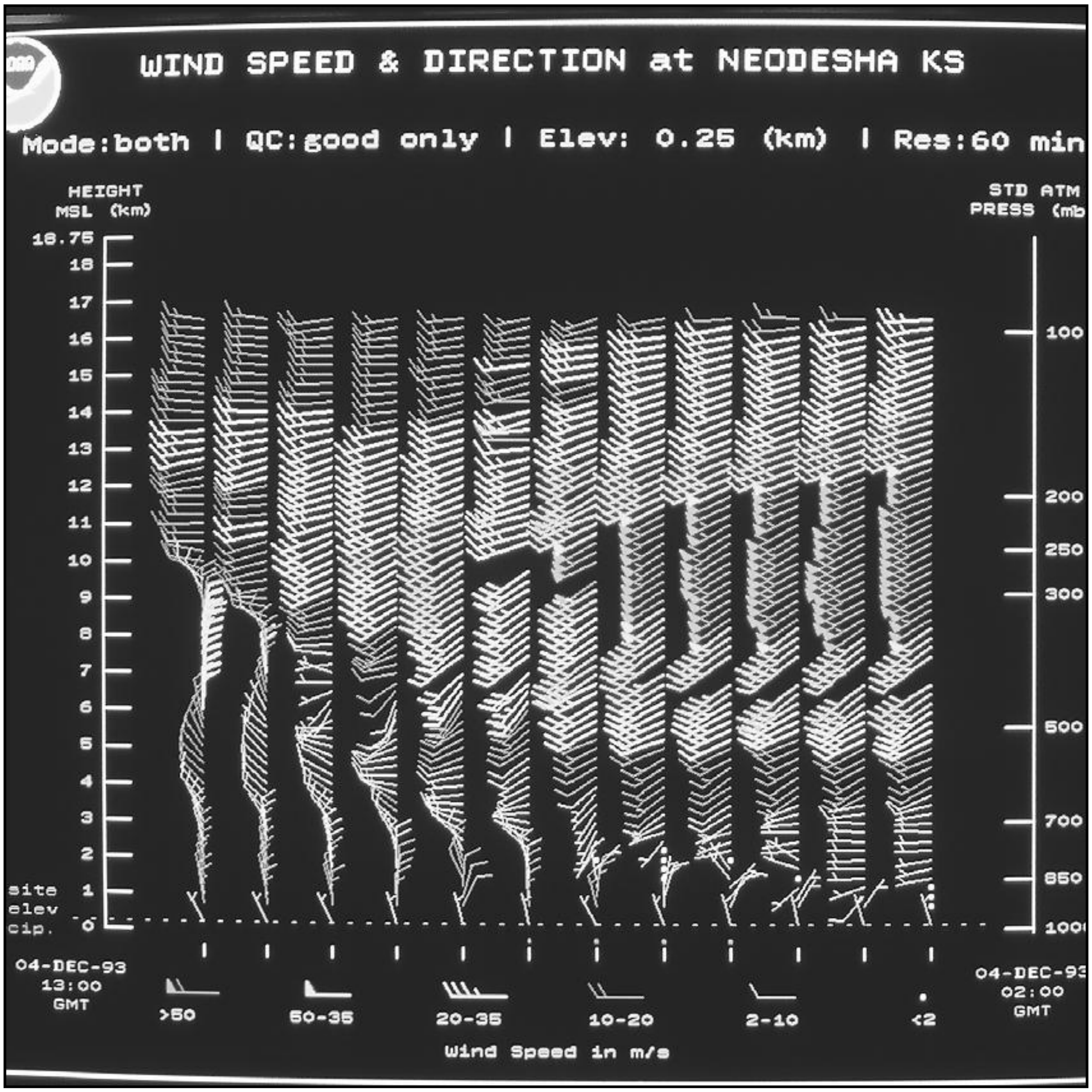
A knowledge of the fundamentals involved in wave motion is critical to understanding the Doppler effect; so, too, is an appreciation of another phenomenon, which is as much related to human psychology and perception as it is to physics. Frame of reference is the perspective of an observer with regard to an object or event. Things may look different for one person in one frame of reference than they do to someone in another.

For example, if you are sitting across the table from a friend at lunch, and you see that he has a spot of ketchup to the right of his mouth, the tendency is to say, "You have some ketchup right here"—and then point to the left of your own mouth, since you are directly across the table from his right. Then he will rub the left side of his face with his napkin, missing the spot entirely, unless you say something like, "No—mirror image." The problem is that each of you has a different frame of reference, yet only your friend took this into account.

RELATIVE MOTION. Physicists often speak of relative motion, or the motion of one object in relation to another. For instance, the molecules in the human body are in a constant state of motion, but they are not moving relative to the body itself: they are moving relative to one another.

On a larger scale, Earth is rotating at a rate of about 1,000 MPH (1,600 km/h), and orbiting the Sun at 67,000 MPH (107,826 km/h)—almost three times as fast as humans have ever traveled in a powered vehicle. Yet no one senses the speed of Earth's movement in the way that one senses the movement of a car—or, indeed, the way the astronauts aboard *Apollo 11* in 1969 perceived that their spacecraft was moving at about 25,000 MPH (40,000 km/h). In the case of the car or the spacecraft, movement can be perceived in relation to other objects: road signs and buildings on the one hand, Earth and the Moon on the other. But humans have no frame of reference from which to perceive the movement of Earth itself.

If one were traveling in a train alongside another train at constant velocity, it would be impossible to perceive that either train was actually moving, unless one looked at a reference point, such as the trees or mountains in the background. Likewise, if two trains were sitting side by side, and one train started to move, the relative motion might cause a passenger in the unmoving train to believe that his or her train



Doppler radar, like the wind speed radar shown here, have become a fundamental tool for meteorologists. *(Photograph by Roger Ressmeyer/Corbis. Reproduced by permission.)*

was the one moving. In fact, as Albert Einstein (1879-1955) demonstrated with his Theory of Relativity, all motion is relative: when we say that something is moving, we mean that it is moving in relation to something else.

Doppler's Discovery

Long before Einstein was born, Austrian physicist Christian Johann Doppler (1803-1853) made an important discovery regarding the relative motion of sound waves or light waves. While teaching in Prague, now the capital of the Czech Republic, but then a part of the Austro-Hungarian Empire, Doppler became fascinated with a common, but previously unexplained, phenomenon. When an observer is standing beside a railroad track and a train approaches, Doppler noticed, the train's whistle has a high pitch. As it passes by, however, the sound of the train whistle suddenly becomes much lower.

By Doppler's time, physicists had recognized the existence of sound waves, as well as the fact a sound's pitch is a function of frequency—in other words, the closer the waves are to one another, the higher the pitch. Taking this knowledge, he reasoned that if a source of sound is moving toward a listener, the waves in front of the source are compressed, thus creating a higher frequency. On the other hand, the waves behind the moving source are stretched out, resulting in a lower frequency.

After developing a mathematical formula to describe this effect, Doppler presented his findings in 1842. Three years later, he and Dutch meteorologist Christopher Heinrich Buys-Ballot (1817-1890) conducted a highly unusual experiment to demonstrate the theory. Buys-Ballot arranged for a band of trumpet players to perform on an open railroad flatcar, while riding past a platform on which a group of musicians with perfect pitch (that is, a finely tuned sense of hearing) sat listening.

The experiment went on for two days, the flatcar passing by again and again, while the horns blasted and the musicians on the platform recorded their observations. Though Doppler and Buys-Ballot must have seemed like crazy men to those who were not involved in the experiment, the result—as interpreted from the musicians' written impressions of the pitches they heard—confirmed Doppler's theory.

REAL-LIFE APPLICATIONS

Sound Compression and the Doppler Effect

As stated in the introduction, one can observe the Doppler effect in a number of settings. If a person is standing by the side of a road and a car approaches at a significant rate of speed, the frequency of the sound waves grows until the car passes the observer, then the frequency suddenly drops. But Doppler, of course, never heard the sound of an automobile, or the siren of a motorized ambulance or fire truck.

In his day, the horse-drawn carriage still constituted the principal means of transportation for short distances, and such vehicles did not attain the speeds necessary for the Doppler effect to become noticeable. Only one mode of transportation in the mid-nineteenth century made it possible to observe and record the effect: a steam-powered locomotive. Therefore, let us consider the Doppler effect as Doppler himself did—in terms of a train passing through a s tation.

THE SOUND OF A TRAIN WHISTLE. When a train is sitting in a station prior to leaving, it blows its whistle, but listeners standing nearby notice nothing unusual. There is no difference—except perhaps in degree of intensity—between the sound heard by someone on the platform, and the sound of the train as heard by someone standing behind the caboose. This is because a stationary train is at the center of the sound waves it produces, which radiate in concentric circles (like a bulls-eye) around it.

As the train begins to move, however, it is no longer at the center of the sound waves emanating from it. Instead, the circle of waves is moving forward, along with the train itself, and, thus, the locomotive compresses waves toward the front. If someone is standing further ahead along the track, that person hears the compressed sound waves. Due to their compression, these have a much higher frequency than the waves produced by a stationary train.

At the same time, someone standing behind the train—a listener on the platform at the station, watching the train recede into the distance—hears the sound waves that emanate from behind the train. It is the same train making the same sound, but because the train has compressed the sound waves in front of it, the waves behind it are spread out, producing a sound of much lower frequency. Thus, the sound of the train, as perceived by two different listeners, varies with frame of reference.

THE SONIC BOOM: A RELATED EFFECT. Some people today have had the experience of hearing a jet fly high overhead, producing a shock wave known as a sonic boom. A sonic boom, needless to say, is certainly not something of which Doppler would have had any knowledge, nor is it an illustration of the Doppler effect, per se. But it is an example of sound compression, and, therefore, it deserves attention here.

The speed of sound, unlike the speed of light, is dependant on the medium through which it travels. Hence, there is no such thing as a fixed "speed of sound"; rather, there is only a speed at which sound waves are transmitted through a given type of material. Its speed through a gas, such as air, is proportional to the square root of the pressure divided by the density. This, in turn, means that the higher the altitude, the slower the speed of sound: for the altitudes at which jets fly, it is about 660 MPH (1,622 km/h).

As a jet moves through the air, it too produces sound waves which compress toward the front, and widen toward the rear. Since sound waves themselves are really just fluctuations in pressure, this means that the faster a jet goes, the greater the pressure of the sound waves bunched up in front of it. Jet pilots speak of "breaking the sound barrier," which is more than just a figure of speech. As the craft approaches the speed of sound, the pilot becomes aware of a wall of high pressure to the front of the plane, and as a result of this high-pressure wall, the jet experiences enormous turbulence.

The speed of sound is referred to as Mach 1, and at a speed of between Mach 1.2 and Mach 1.4, even stranger things begin to happen. Now the jet is moving faster than the sound waves emanating from it, and, therefore, an observer on the ground sees the jet move by well before hearing the sound. Of course, this would happen to some extent anyway, since light travels so much faster than sound; but the difference between the arrival time of the light waves and the sound waves is even more noticeable in this situation.

Meanwhile, up in the air, every protruding surface of the aircraft experiences intense pressure: in particular, sound waves tend to become highly compressed along the aircraft's nose and tail. Eventually these compressed sound waves build up, resulting in a shock wave. Down on the ground, the shock wave manifests as a "sonic boom"—or rather, two sonic booms—one from the nose of the craft, and one from the tail. People in the aircraft do not hear the boom, but the shock waves produced by the compressed sound can cause sudden changes in pressure, density, and temperature that can pose dangers to the operation of the airplane. To overcome this problem, designers of supersonic aircraft have developed planes with wings that are swept back, so they fit within the cone of pressure.

Doppler Radar and Other Sensing Technology

The Doppler effect has a number of applications relating to the sensing of movement. For instance, physicians and medical technicians apply it to measure the rate and direction of blood flow in a patient's body, along with ultrasound. As blood moves through an artery, its top speed is 0.89 MPH (0.4 m/s)—not very fast, yet fast enough, given the small area in which movement is taking place, for the Doppler effect to be observed. A beam of ultrasound is pointed toward an artery, and the reflected waves exhibit a shift in frequency, because the blood cells are acting as moving sources of sound waves—just like the trains Doppler observed.

Not all applications of the Doppler effect fall under the heading of "technology": some can be found in nature. Bats use the Doppler effect to hunt for prey. As a bat flies, it navigates by emitting whistles and listening for the echoes. When it is chasing down food, its brain detects a change in pitch between the emitted whistle, and the echo it receives. This tells the bat the speed of its quarry, and the bat adjusts its own speed accordingly.

DOPPLER RADAR. Police officers may not enjoy the comparison—given the public's general impression of bats as evil, bloodthirsty creatures—but in using radar as a basis to check for speeding violations, the police are applying a principle similar to that used by bats. Doppler radar, which uses the Doppler effect to calculate the speed of moving objects, is a form of technology used not only by law-enforcement officers, but also by meteorologists.

The change in frequency experienced as a result of the Doppler effect is exactly twice the ratio between the velocity of the target (for instance, a speeding car) and the speed with which the radar pulse is directed toward the target. From this formula, it is possible to determine the velocity of the target when the frequency change and speed of radar propagation are known. The police officer's Doppler radar performs these calculations; then all the officer has to do is pull over the speeder and write a ticket.

Meteorologists use Doppler radar to track the movement of storm systems. By detecting the direction and velocity of raindrops or hail, for instance, Doppler radar can be used to determine the motion of winds and, thus, to predict weather patterns that will follow in the next minutes or hours. But Doppler radar can do more than simply detect a storm in progress: Doppler technology also aids meteorologists by interpreting wind direction, as an indicator of coming storms.

The Doppler Effect in Light Waves

So far the Doppler effect has been discussed purely in terms of sound waves; but Doppler

KEY TERMS

AMPLITUDE: The maximum displacement of a vibrating material. In wave motion, amplitude is the "size" of a wave, an indicator of the energy and intensity of the wave.

CYCLE: One complete oscillation. In wave motion, this is equivalent to the movement of a wave from trough to crest and back to trough. For a sound wave, in particular, a cycle is one complete vibration.

DOPPLER EFFECT: The change in the observed frequency of a wave when the source of the wave is moving with respect to the observer. It is named after Austrian physicist Johann Christian Doppler (1803-1853), who discovered it.

FRAME OF REFERENCE: The perspective an observer has with regard to an object or action. Frame of reference affects perception of various physical properties, and plays a significant role in the Doppler effect.

FREQUENCY: In wave motion, frequency is the number of waves passing through a given point during the interval of one second. The higher the frequency, the shorter the wavelength. Measured in Hertz, frequency is mathematically related to wave speed, wavelength, and period.

HARMONIC MOTION: The repeated movement of a particle about a position of equilibrium, or balance.

HERTZ: A unit for measuring frequency, named after nineteenth-century German physicist Heinrich Rudolf Hertz (1857-1894). High frequencies are expressed in terms of kilohertz (kHz; 10^3 or 1,000 cycles per second); megahertz (MHz; 10^6 or 1 million cycles per second); and gigahertz (GHz; 10^9 or 1 billion cycles per second.)

INTENSITY: Intensity is the rate at which a wave moves energy per unit of cross-sectional area. Where sound waves are concerned, intensity is commonly known as "volume."

OSCILLATION: A type of harmonic motion, typically periodic, in one or more dimensions.

PERIOD: For wave motion, a period is the amount of time required to complete one full cycle. Period is mathematically related to frequency, wavelength, and wave speed.

PERIODIC MOTION: Motion that is repeated at regular intervals. These intervals are known as periods.

PERIODIC WAVE: A wave in which a uniform series of crests and troughs follow one after the other in regular succession.

RELATIVE MOTION: The motion of one object in relation to another.

WAVELENGTH: The distance between a crest and the adjacent crest, or the trough and an adjacent trough, of a wave. Wavelength, symbolized λ (the Greek letter lambda) is mathematically related to wave speed, period, and frequency.

WAVE MOTION: A type of harmonic motion that carries energy from one place to another without actually moving any matter.

himself maintained that it could be applied to light waves as well, and experimentation conducted in 1901 proved him correct. This was far from an obvious point, since light is quite different from sound.

Not only does light travel much, much faster—186,000 mi (299,339 km) a second—but unlike sound, light does not need to travel through a medium. Whereas sound cannot be transmitted in outer space, light is transmitted by radiation, a form of energy transfer that can be directed as easily through a vacuum as through matter.

The Doppler effect in light can be demonstrated by using a device called a spectroscope, which measures the spectral lines from an object of known chemical composition. These spectral lines are produced either by the absorption or emission of specific frequencies of light by electrons in the source material. If the light waves appear at the blue, or high-frequency end of the visible light spectrum, this means that the object is moving toward the observer. If, on the other hand, the light waves appear at the red, or low-frequency end of the spectrum, the object is moving away.

HUBBLE AND THE RED SHIFT. In 1923, American astronomer Edwin Hubble (1889-1953) observed that the light waves from distant galaxies were shifted so much to the red end of the light spectrum that they must be moving away from the Milky Way, the galaxy in which Earth is located, at a high rate. At the same time, nearer galaxies experienced much less of a red shift, as this phenomenon came to be known, meaning that they were moving away at relatively slower speeds.

Six years later, Hubble and another astronomer, Milton Humason, developed a mathematical formula whereby astronomers could determine the distance to another galaxy by measuring that galaxy's red shifts. The formula came to be known as Hubble's constant, and it established the relationship between red shift and the velocity at which a galaxy or object was receding from Earth. From Hubble's work, it became clear that the universe was expanding, and research by a number of physicists and astronomers led to the development of the "big bang" theory—the idea that the universe emerged almost instantaneously, in some sort of explosion, from a compressed state of matter.

WHERE TO LEARN MORE

Beiser, Arthur. *Physics,* 5th ed. Reading, MA: Addison-Wesley, 1991.

Bryant-Mole, Karen. *Sound and Light.* Crystal Lake, IL: Rigby Interactive Library, 1997.

Challoner, Jack. *Sound and Light.* New York: Kingfisher, 2001.

Dispenzio, Michael A. *Awesome Experiments in Light and Sound.* Illustrated by Catherine Leary. New York: Sterling Publishing Company, 1999.

"The Doppler Effect." The Physics Classroom (Web site). <http://www.glenbrook.k12.il.us/gbssci/phys/Class/waves/u10l3d.html> (April 29, 2001).

Maton, Anthea. *Exploring Physical Science.* Upper Saddle River, N.J.: Prentice Hall, 1997.

Russell, David A. *"The Doppler Effect and Sonic Booms" Kettering University* (Web site). <http://www.kettering.edu/~drussell/Demos/doppler/doppler.html> (April 29, 2001).

Snedden, Robert. *Light and Sound.* Des Plaines, IL: Heinemann Library, 1999.

"Sound Wave—Doppler Effect" (Web site). <http://csgrad.cs.vt.edu/~chin/doppler.html> (April 29, 2001).

"Wave Motion—Doppler Effect" (Web site). <http://members.aol.com/cepeirce/b21.html> (April 29, 2001).

SOUND

ACOUSTICS

ULTRASONICS

ACOUSTICS

CONCEPT

The area of physics known as acoustics is devoted to the study of the production, transmission, and reception of sound. Thus, wherever sound is produced and transmitted, it will have an effect somewhere, even if there is no one present to hear it. The medium of sound transmission is an all-important, key factor. Among the areas addressed within the realm of acoustics are the production of sounds by the human voice and various instruments, as well as the reception of sound waves by the human ear.

HOW IT WORKS

Wave Motion and Sound waves

Sound waves are an example of a larger phenomenon known as wave motion, and wave motion is, in turn, a subset of harmonic motion—that is, repeated movement of a particle about a position of equilibrium, or balance. In the case of sound, the "particle" is not an item of matter, but of energy, and wave motion is a type of harmonic movement that carries energy from one place to another without actually moving any matter.

Particles in waves experience oscillation, harmonic motion in one or more dimensions. Oscillation itself involves little movement, though some particles do move short distances as they interact with other particles. Primarily, however, it involves only movement in place. The waves themselves, on the other hand, move across space, ending up in a position different from the one in which they started.

A transverse wave forms a regular up-and-down pattern in which the oscillation is perpendicular to the direction the wave is moving. This is a fairly easy type of wave to visualize: imagine a curve moving up and down along a straight line. Sound waves, on the other hand, are longitudinal waves, in which oscillation occurs in the same direction as the wave itself.

These oscillations are really just fluctuations in pressure. As a sound wave moves through a medium such as air, these changes in pressure cause the medium to experience alternations of density and rarefaction (a decrease in density). This, in turn, produces vibrations in the human ear or in any other object that receives the sound waves.

Properties of Sound Waves

CYCLE AND PERIOD. The term cycle has a definition that varies slightly, depending on whether the type of motion being discussed is oscillation, the movement of transverse waves, or the motion of a longitudinal sound wave. In the latter case, a cycle is defined as a single complete vibration.

A period (represented by the symbol *T*) is the amount of time required to complete one full cycle. The period of a sound wave can be mathematically related to several other aspects of wave motion, including wave speed, frequency, and wavelength.

THE SPEED OF SOUND IN VARIOUS MEDIA. People often refer to the "speed of sound" as though this were a fixed value like the speed of light, but, in fact, the speed of sound is a function of the medium through which it travels. What people ordinarily mean by the "speed of sound" is the speed of sound through air at a specific temperature. For sound

BECAUSE THE SOUND GENERATED BY A JET ENGINE CAN DAMAGE A PERSON'S HEARING, AIRPORT GROUND CREWS ALWAYS WEAR PROTECTIVE HEADGEAR. *(Photograph by Patrick Bennett/Corbis. Reproduced by permission.)*

traveling at sea level, the speed at 32°F (0°C) is 740 MPH (331 m/s), and at 68°F (20°C), it is 767 MPH (343 m/s).

In the essay on aerodynamics, the speed of sound for aircraft was given at 660 MPH (451 m/s). This is much less than the figures given above for the speed of sound through air at sea level, because obviously, aircraft are not flying at sea level, but well above it, and the air through which they pass is well below freezing temperature.

The speed of sound through a gas is proportional to the square root of the pressure divided by the density. According to Gay-Lussac's law, pressure is directly related to temperature, meaning that the lower the pressure, the lower the temperature—and vice versa. At high altitudes, the temperature is low, and, therefore, so is the pressure; and, due to the relatively small gravitational pull that Earth exerts on the air at that height, the density is also low. Hence, the speed of sound is also low.

It follows that the higher the pressure of the material, and the greater the density, the faster sound travels through it: thus sound travels faster through a liquid than through a gas. This might seem a bit surprising: at first glance, it would seem that sound travels fastest through air, but only because we are just more accustomed to hearing sounds that travel through that medium. The speed of sound in water varies from about 3,244 MPH (1,450 m/s) to about 3,355 MPH (1500 m/s). Sound travels even faster through a solid—typically about 11,185 MPH (5,000 m/s)—than it does through a liquid.

FREQUENCY. Frequency (abbreviated *f*) is the number of waves passing through a given point during the interval of one second. It is measured in Hertz (Hz), named after nineteenth-century German physicist Heinrich Rudolf Hertz (1857-1894) and a Hertz is equal to one cycle of oscillation per second. Higher frequencies are expressed in terms of kilohertz (kHz; 10^3 or 1,000 cycles per second) or megahertz (MHz; 10^6 or 1 million cycles per second.)

The human ear is capable of hearing sounds from 20 to approximately 20,000 Hz—a relatively small range for a mammal, considering that bats, whales, and dolphins can hear sounds at a frequency up to 150 kHz. Human speech is in the range of about 1 kHz, and the 88 keys on a piano vary in frequency from 27 Hz to 4,186 Hz. Each note has its own frequency, with middle C (the "white key" in the very middle of a piano keyboard) at 264 Hz. The quality of harmony or dissonance when two notes are played together is a

PIANO STRINGS GENERATE SOUND AS THEY ARE SET INTO VIBRATION BY THE HAMMERS. THE HAMMERS, IN TURN, ARE ATTACHED TO THE BLACK-AND-WHITE KEYS ON THE OUTSIDE OF THE PIANO. *(Photograph by Bob Krist/Corbis. Reproduced by permission.)*

function of the relationship between the frequencies of the two.

Frequencies below the range of human audibility are called infrasound, and those above it are referred to as ultrasound. There are a number of practical applications for ultrasonic technology in medicine, navigation, and other fields.

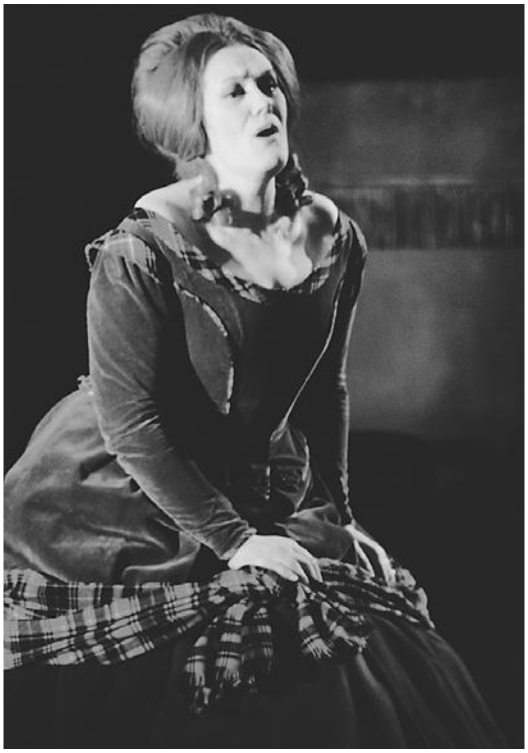
WAVELENGTH. Wavelength (represented by the symbol λ, the Greek letter lambda) is the distance between a crest and the adjacent crest, or a trough and an adjacent trough, of a wave. The higher the frequency, the shorter the wavelength, and vice versa. Thus, a frequency of 20 Hz, at the bottom end of human audibility, has a very large wavelength: 56 ft (17 m). The top end frequency of 20,000 Hz is only 0.67 inches (17 mm).

There is a special type of high-frequency sound wave beyond ultrasound: hypersound, which has frequencies above 10^7 MHz, or 10 trillion Hz. It is almost impossible for hypersound waves to travel through all but the densest media, because their wavelengths are so short. In order to be transmitted properly, hypersound requires an extremely tight molecular structure; otherwise, the wave would get lost between molecules.

Wavelengths of visible light, part of the electromagnetic spectrum, have a frequency much higher even than hypersound waves: about 10^9 MHz, 100 times greater than for hypersound. This, in turn, means that these wavelengths are incredibly small, and this is why light waves can easily be blocked out by using one's hand or a curtain.

The same does not hold for sound waves, because the wavelengths of sounds in the range of human audibility are comparable to the size of ordinary objects. To block out a sound wave, one needs something of much greater dimensions—width, height, and depth—than a mere cloth curtain. A thick concrete wall, for instance, may be enough to block out the waves. Better still would be the use of materials that absorb sound, such as cork, or even the use of machines that produce sound waves which destructively interfere with the offending sound.

AMPLITUDE AND INTENSITY. Amplitude is critical to the understanding of sound, though it is mathematically independent from the parameters so far discussed. Defined as the maximum displacement of a vibrating material, amplitude is the "size" of a wave. The greater the amplitude, the greater the energy the wave



The human voice, whether in speech or in song, is a remarkable sound-producing instrument: at any given moment as a person is talking or singing, parts of the vocal cords are opened, and parts are closed. Shown here is opera superstar Joan Sutherland. *(Hulton-Deutsch Collection/Corbis. Reproduced by permission.)*

contains: amplitude indicates intensity, commonly known as "volume," which is the rate at which a wave moves energy per unit of a cross-sectional area.

Intensity can be measured in watts per square meter, or W/m^2. A sound wave of minimum intensity for human audibility would have a value of 10^{-12}, or 0.000000000001, W/m^2. As a basis of comparison, a person speaking in an ordinary tone of voice generates about 10^{-4}, or 0.0001, watts. On the other hand, a sound with an intensity of 1 W/m^2 would be powerful enough to damage a person's ears.

REAL-LIFE APPLICATIONS

Decibel Levels

For measuring the intensity of a sound as experienced by the human ear, we use a unit other than the watt per square meter, because ears do not respond to sounds in a linear, or straight-line, progression. If the intensity of a sound is doubled, a person perceives a greater intensity, but nothing approaching twice that of the original sound. Instead, a different system—known in mathematics as a logarithmic scale—is applied.

In measuring the effect of sound intensity on the human ear, a unit called the decibel (abbreviated dB) is used. A sound of minimal audibility (10^{-12} W/m^2) is assigned the value of 0 dB, and 10 dB is 10 times as great—10^{-11} W/m^2. But 20 dB is not 20 times as intense as 0 dB; it is 100 times as intense, or 10^{-10} W/m^2. Every increase of 10 dB thus indicates a tenfold increase in intensity. Therefore, 120 dB, the maximum decibel level that a human ear can endure without experiencing damage, is not 120 times as great as the minimal level for audibility, but 10^{12} (1 trillion) times as great—equal to 1 W/m^2, referred to above as the highest safe intensity level.

Of course, sounds can be much louder than 120 dB: a rock band, for instance, can generate sounds of 125 dB, which is 5 times the maximum safe decibel level. A gunshot, firecracker, or a jet—if one is exposed to these sounds at a sufficiently close proximity—can be as high as 140 dB, or 20 times the maximum safe level. Nor is 120 dB safe for prolonged periods: hearing experts indicate that regular and repeated exposure to even 85 dB (5 less than a lawn mower) can cause permanent damage to one's hearing.

Production of Sound Waves

MUSICAL INSTRUMENTS. Sound waves are vibrations; thus, in order to produce sound, vibrations must be produced. For a stringed instrument, such as a guitar, harp, or piano, the strings must be set into vibration, either by the musician's fingers or the mechanism that connects piano keys to the strings inside the case of the piano.

In other woodwind instruments and horns, the musician causes vibrations by blowing into the mouthpiece. The exact process by which the vibrations emerge as sound differs between woodwind instruments, such as a clarinet or saxophone on the one hand, and brass instruments, such as a trumpet or trombone on the other. Then there is a drum or other percussion instrument, which produces vibrations, if not musical notes.

ELECTRONIC AMPLIFICATION. Sound is a form of energy: thus, when an automobile or other machine produces sound incidental to its operation, this actually represents energy that is lost. Energy itself is conserved, but not all of the energy put into the machine can ever be realized as useful energy; thus, the automobile loses some energy in the form of sound and heat.

The fact that sound is energy, however, also means that it can be converted to other forms of energy, and this is precisely what a microphone does: it receives sound waves and converts them to electrical energy. These electrical signals are transmitted to an amplifier, and next to a loudspeaker, which turns electrical energy back into sound energy—only now, the intensity of the sound is much greater.

Inside a loudspeaker is a diaphragm, a thin, flexible disk that vibrates with the intensity of the sound it produces. When it pushes outward, the diaphragm forces nearby air molecules closer together, creating a high-pressure region around the loudspeaker. (Remember, as stated earlier, that sound is a matter of fluctuations in pressure.) The diaphragm is then pushed backward in response, freeing up an area of space for the air molecules. These, then, rush toward the diaphragm, creating a low-pressure region behind the high-pressure one. The loudspeaker thus sends out alternating waves of high and low pressure, vibrations on the same frequency of the original sound.

THE HUMAN VOICE. As impressive as the electronic means of sound production are (and of course the description just given is highly simplified), this technology pales in comparison to the greatest of all sound-producing mechanisms: the human voice. Speech itself is a highly complex physical process, much too involved to be discussed in any depth here. For our present purpose, it is important only to recognize that speech is essentially a matter of producing vibrations on the vocal cords, and then transmitting those vibrations.

Before a person speaks, the brain sends signals to the vocal cords, causing them to tighten. As speech begins, air is forced across the vocal cords, and this produces vibrations. The action of the vocal cords in producing these vibrations is, like everything about the miracle of speech, exceedingly involved: at any given moment as a person is talking, parts of the vocal cords are opened, and parts are closed.

The sound of a person's voice is affected by a number of factors: the size and shape of the sinuses and other cavities in the head, the shape of the mouth, and the placement of the teeth and tongue. These factors influence the production of specific frequencies of sound, and result in differing vocal qualities. Again, the mechanisms of speech are highly complicated, involving action of the diaphragm (a partition of muscle and tissue between the chest and abdominal cavities), larynx, pharynx, glottis, hard and soft palates, and so on. But, it all begins with the production of vibrations.

Propagation: Does It Make a Sound?

As stated in the introduction, acoustics is concerned with the production, transmission (sometimes called propagation), and reception of sound. Transmission has already been examined in terms of the speed at which sound travels through various media. One aspect of sound transmission needs to be reiterated, however: for sound to be propagated, there must be a medium.

There is an age-old "philosophical" question that goes something like this: If a tree falls in the woods and there is no one to hear it, does it make a sound? In fact, the question is not a matter of philosophy at all, but of physics, and the answer is, of course, "yes." As the tree falls, it releases energy in a number of forms, and part of this energy is manifested as sound waves.

Consider, on the other hand, this rephrased version of the question: "If a tree falls in a vacuum—an area completely devoid of matter, including air—does it make a sound?" The answer is now a qualified "no": certainly, there is a release of energy, as before, but the sound waves cannot be transmitted. Without air or any other matter to carry the waves, there is literally no sound.

Hence, there is a great deal of truth to the tagline associated with the 1979 science-fiction film *Alien*: "In space, no one can hear you scream." Inside an astronaut's suit, there is pressure and an oxygen supply; without either, the astronaut would perish quickly. The pressure and air inside the suit also allow the astronaut to hear sounds within the suit, including communica-

KEY TERMS

ACOUSTICS: An area of physics devoted to the study of the production, transmission, and reception of sound.

AMPLITUDE: The maximum displacement of a vibrating material. In wave motion, amplitude is the "size" of a wave, and for sound waves, amplitude indicates the intensity or volume of sound.

CYCLE: For a sound wave, a cycle is a single complete vibration.

DECIBEL: A unit for measuring intensity of sound. Decibels, abbreviated dB, are calibrated along a logarithmic scale whereby every increase of 10 dB indicates an increase in intensity by a factor of 10. Thus if the level of intensity is increased from 30 to 60 dB, the resulting intensity is not twice as great as that of the earlier sound—it is 1,000 times as great.

ENERGY: The ability to perform work, which is the exertion of force over a given distance. Work is the product of force and distance, where force and distance are exerted in the same direction.

FREQUENCY: In wave motion, frequency is the number of waves passing through a given point during the interval of one second. The higher the frequency, the shorter the wavelength. Measured in Hertz, frequency is mathematically related to wave speed, wavelength, and period.

HARMONIC MOTION: The repeated movement of a particle about a position of equilibrium, or balance.

HERTZ: A unit for measuring frequency, named after nineteenth-century German physicist Heinrich Rudolf Hertz (1857-1894). High frequencies are expressed in terms of kilohertz (kHz; 10^3 or 1,000 cycles per second) or megahertz (MHz; 10^6 or 1 million cycles per second.)

INTENSITY: Intensity is the rate at which a wave moves energy per unit of cross-sectional area. Where sound waves are concerned, intensity is commonly known as "volume."

tions via microphone from other astronauts. But, if there were an explosion in the vacuum of deep space outside the spacecraft, no one inside would be able to hear it.

Reception of Sound

RECORDING. Earlier the structure of electronic amplification was described in very simple terms. Some of the same processes—specifically, the conversion of sound to electrical energy—are used in the recording of sound. In sound recording, when a sound wave is emitted, it causes vibrations in a diaphragm attached to an electrical condenser. This causes variations in the electrical current passed on by the condenser.

These electrical pulses are processed and ultimately passed on to an electromagnetic "recording head." The magnetic field of the recording head extends over the section of tape being recorded: what began as loud sounds now produce strong magnetic fields, and soft sounds produce weak fields. Yet, just as electronic means of sound production and transmission are still not as impressive as the mechanisms of the human voice, so electronic sound reception and recording technology is a less magnificent device than the human ear.

HOW THE EAR HEARS. As almost everyone has noticed, a change in altitude (and, hence, of atmospheric pressure) leads to a

KEY TERMS CONTINUED

LONGITUDINAL WAVE: A wave in which the movement of vibration is in the same direction as the wave itself. A sound wave is an example of a longitudinal wave.

MATTER: Physical substance that has mass; occupies space; is composed of atoms; and is ultimately convertible to energy.

MEDIUM: Material through which sound travels. (It cannot travel through a vacuum.) The most common medium (plural, media) of sound transmission experienced in daily life is air, but in fact sound can travel through any type of matter.

OSCILLATION: The vibration experienced by individual waves even as the wave itself is moving through space. Oscillation is a type of harmonic motion, typically periodic, in one or more dimensions.

PERIOD: For wave motion, a period is the amount of time required to complete one full cycle. Period is mathematically related to frequency, wavelength, and wave speed.

PERIODIC MOTION: Motion that is repeated at regular intervals. These intervals are known as periods.

RAREFACTION: A decrease in density.

ULTRASOUND : Sound waves with a frequency above 20,000 Hertz, which makes them inaudible to the human ear.

VACUUM: An area entirely devoid of matter, including air.

WAVELENGTH: The distance between a crest and the adjacent crest, or the trough and an adjacent trough, of a wave. Wavelength, symbolized by λ (the Greek letter lambda) is mathematically related to wave speed, period, and frequency.

WAVE MOTION: A type of harmonic motion that carries energy from one place to another without actually moving any matter.

strange "popping" sensation in the ears. Usually, this condition can be overcome by swallowing, or even better, by yawning. This opens the Eustachian tube, a passageway that maintains atmospheric pressure in the ear. Useful as it is, the Eustachian tube is just one of the human ear's many parts.

The "funny" shape of the ear helps it to capture and amplify sound waves, which pass through the ear canal and cause the eardrum to vibrate. Though humans can hear sounds over a much wider range, the optimal range of audibility is from 3,000 to 4,000 Hz. This is because the structure of the ear canal is such that sounds in this frequency produce magnified pressure fluctuations. Thanks to this, as well as other specific properties, the ear acts as an amplifier of sounds.

Beyond the eardrum is the middle ear, an intricate sound-reception device containing some of the smallest bones in the human body—bones commonly known, because of their shapes, as the hammer, anvil, and stirrup. Vibrations pass from the hammer to the anvil to the stirrup, through the membrane that covers the oval window, and into the inner ear.

Filled with liquid, the inner ear contains the semicircular canals responsible for providing a sense of balance or orientation: without these, a person literally "would not know which way is up." Also, in the inner ear is the cochlea, an organ shaped like a snail. Waves of pressure from the fluids of the inner ear are passed through the cochlea to the auditory nerve, which then transmits these signals to the brain.

The basilar membrane of the cochlea is a particularly wondrous instrument, responsible in large part for the ability to discriminate between sounds of different frequencies and intensities. The surface of the membrane is covered with thousands of fibers, which are highly sensitive to disturbances, and it transmits information concerning these disturbances to the auditory nerve. The brain, in turn, forms a relation between the position of the nerve ending and the frequency of the sound. It also equates the degree of disturbance in the basilar membrane with the intensity of the sound: the greater the disturbance, the louder the sound.

WHERE TO LEARN MORE

Adams, Richard C. and Peter H. Goodwin. *Engineering Projects for Young Scientists.* New York: Franklin Watts, 2000.

Beiser, Arthur. *Physics*, 5th ed. Reading, MA: Addison-Wesley, 1991.

Friedhoffer, Robert. *Sound.* Illustrated by Richard Kaufman and Linda Eisenberg; photographs by Timothy White. New York: F. Watts, 1992.

Gardner, Robert. *Science Projects About Sound.* Berkeley Heights, NJ: Enslow Publishers, 2000.

Internet Resources for Sound and Light (Web site). <http://electro.sau.edu/SLResources.html> (April 25, 2001).

"*Music and Sound Waves*" (Web site). <http://www.silcom.com/~aludwig/musicand.htm> (April 28, 2001).

Oxlade, Chris. *Light and Sound.* Des Plaines, IL: Heinemann Library, 2000.

Physics Tutorial System: Sound Waves Modules (Web site). <http://csgrad.cs.vt.edu/~chin/chin_sound.html> (April 25, 2001).

"*Sound Waves and Music.*" *The Physics Classroom* (Web site). <http://www.glenbrook.k12.il.us/gbssci/phys/Class/sound/soundtoc.h tml> (April 28, 2001).

"*What Are Sound Waves?*" (Web site). <http://rustam.uwp.edu/GWWM/sound_waves.html> (April 28, 2001).

ULTRASONICS

CONCEPT

The word ultrasonic combines the Latin roots ultra, meaning "beyond," and sonic, or sound. The field of ultrasonics thus involves the use of sound waves outside the audible range for humans. These sounds have applications for imaging, detection, and navigation—from helping prospective parents get a glimpse of their unborn child to guiding submarines through the oceans. Ultrasonics can be used to join materials, as for instance in welding or the homogenization of milk, or to separate them, as for example in extremely delicate cleaning operations. Among the broad sectors of society that regularly apply ultrasonic technology are the medical community, industry, the military, and private citizens.

HOW IT WORKS

In the realm of physics, ultrasonics falls under the category of studies in sound. Sound itself fits within the larger heading of wave motion, which is in turn closely related to vibration, or harmonic (back-and-forth) motion. Both wave motion and vibration involve the regular repetition of a certain form of movement; and in both, potential energy (think of the energy in a sled at the top of a hill) is continually converted to kinetic energy (like the energy of a sled as it is sliding down the hill) and back again.

Wave motion carries energy from one place to another without actually moving any matter. Waves themselves may consist of matter, as for instance in the case of a wave on a plucked string or the waves on the ocean. This type of wave is called a mechanical wave, but again, the matter itself does not undergo any net displacement over horizontal space: contrary to what our eyes tell us, molecules of water in an ocean wave move up and down, but they do not actually travel with the wave itself. Only the energy is moved.

SOUND WAVES

Then there are waves of pulses, such as light, sound, radio, or electromagnetic waves. Sound travels by means of periodic waves, a period being the amount of time it takes a complete wave, from trough to crest and back again, to pass through a given point. These periodic waves are typified by a sinusoidal pattern. To picture a sinusoidal wave, one need only imagine an x-axis crossed at regular intervals by a curve that rises above the line to point y before moving downward, below the axis, to point y. This may be expressed also as a graph of sin x versus x. In any case, the wave varies by equal distances upward and downward as it moves along the x-axis in a regular, unvarying pattern.

Periodic waves have three notable interrelated characteristics. One of these is speed, typically calculated in seconds. Another is wavelength, or the distance between a crest and the adjacent crest, or a trough and the adjacent trough, along a plane parallel to that of the wave itself. Finally, there is frequency, the number of waves passing through a given point during the interval of one second.

Frequency is measured in terms of cycles per second, or Hertz (Hz), named in honor of the nineteenth-century German physicist Heinrich Hertz. If a wave has a frequency of 100 Hz, this means that 100 waves are passing through a given point during the interval of one second. Higher frequencies are expressed in terms of kilohertz

Submarines, such as the U.S. Navy submarine pictured here, rely heavily on sonar. *(Corbis. Reproduced by permission.)*

(kHz; 10^3 or 1,000 cycles per second) or megahertz (MHz; 10^6 or 1 million cycles per second.)

Clearly, frequency is a function of the wave's speed or velocity, and the same relationship—though it is not so obvious intuitively—exists between wavelength and speed. Over the interval of one second, a given number of waves pass a certain point (frequency), and each wave occupies a certain distance (wavelength). Multiplied by one another, these two properties equal the velocity of the wave.

An additional characteristic of waves (though one that is not related mathematically to the three named above) is amplitude, or maximum displacement, which can be described as the distance from the x-axis to either the crest or the trough. Amplitude is related to the intensity or the amount of energy in the wave.

These four qualities are easiest to imagine on a transverse wave, described earlier with reference to the x-axis—a wave, in other words, in which vibration or harmonic motion occurs perpendicular to the direction in which the wave is moving.

Such a wave is much easier to picture, for the purposes of illustrating concepts such as frequency, than a longitudinal wave; but in fact, sound waves are longitudinal. A longitudinal wave is one in which the individual segments vibrate in the same direction as the wave itself. The shock waves of an explosion, or the concentric waves of a radio transmission as it goes out from the station to all points within receiving distance, are examples of longitudinal waves. In this type of wave pattern, the crests and troughs are not side by side in a line; they radiate outward. Wavelength is the distance between each concentric circle or semicircle (that is, wave), and amplitude the "width" of each wave, which one may imagine by likening it to the relative width of colors on a rainbow.

Having identified its shape, it is reasonable to ask what, exactly, a sound wave is. Simply put, sound waves are changes in pressure, or an alternation between condensation and rarefaction. Imagine a set of longitudinal waves—represented as concentric circles—radiating from a sound source. The waves themselves are relatively higher in pressure, or denser, than the "spaces" between them, though this is just an illustration for the purposes of clarity: in fact the "spaces" are waves of lower pressure that alternate with higher-pressure waves.

Vibration is integral to the generation of sound. When the diaphragm of a loudspeaker

A fishing revolution? It's possible, thanks to this electronic fish-finder developed by Matsushita. The device includes a detachable sonar sensor. *(AFP/Corbis. Reproduced by permission.)*

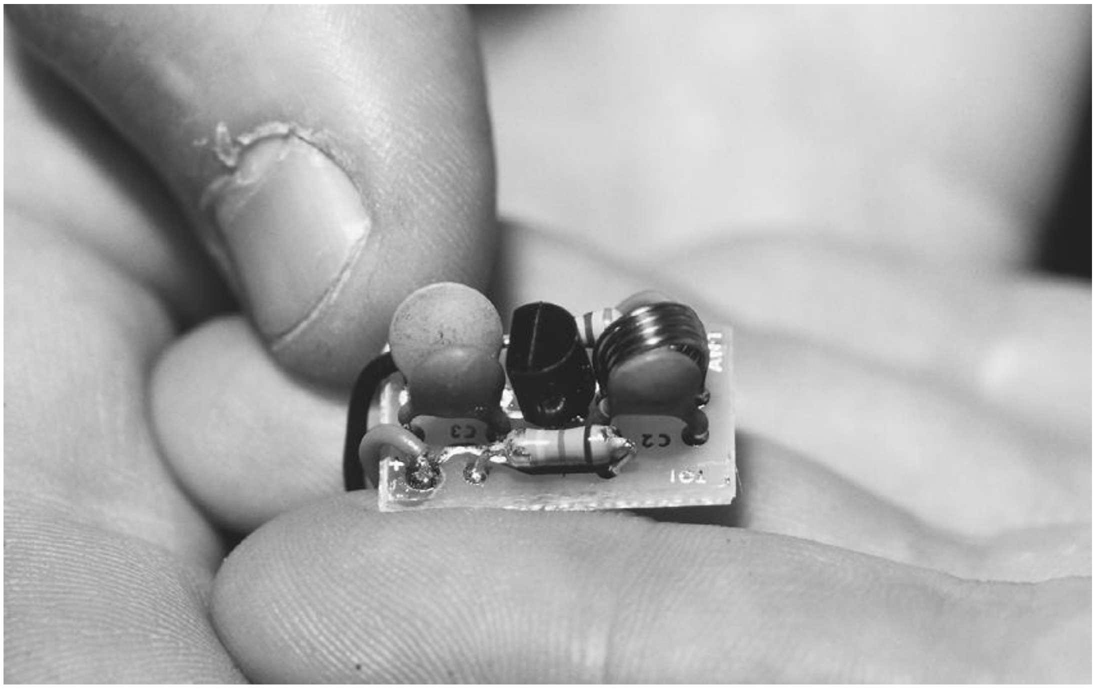
pushes outward, it forces nearby air molecules closer together, creating a high-pressure region all around the loudspeaker. The diaphragm is pushed backward in response, thus freeing up a volume of space for the air molecules. These then rush toward the diaphragm, creating a low-pressure region behind the high-pressure one. As a result, the loudspeaker sends out alternating waves of high pressure (condensation) and low pressure (rarefaction). Furthermore, as sound waves pass through a medium—air, for the purposes of this discussion—they create fluctuations between density and rarefaction. These result in pressure changes that cause the listener's eardrum to vibrate with the same frequency as the sound wave, a vibration that the ear's inner mechanisms translate and pass on to the brain.

The Speed of Sound: Consider the Medium

The speed of sound varies with the hardness of the medium through which it passes: contrary to what you might imagine, it travels faster through liquids than through gasses such as air, and faster through solids than through liquids. By definition, molecules are closer together in harder material, and thus more quickly responsive to signals from neighboring particles. In granite, for instance, sound travels at 19,680 ft per second (6,000 mps), whereas in air, the speed of sound is only 1,086 ft per second (331 mps). It follows that sound travels faster in water—5,023 ft per second (1,531 mps), to be exact—than in air. It should be clear, then, that there is a correlation between density and the ease with which a sound travels. Thus, sound cannot travel in a vacuum, giving credence to the famous tagline from the 1979 science fiction thriller *Alien:* "In space, no one can hear you scream."

When sound travels through a medium such as air, however, two factors govern its audibility: intensity or volume (related to amplitude and measured in decibels, or dB) and frequency. There is no direct correlation between intensity and frequency, though for a person to hear a very low-frequency sound, it must be above a certain decibel level. (At all frequencies, however, the threshold of discomfort is around 120 decibels.)

In any case, when discussing ultrasonics, frequency and not intensity is of principal concern. The range of audibility for the human ear is from 20 Hz to 20,000 Hz, with frequencies below that range dubbed infrasound and those above it referred to as ultrasound. (There is a third category, hypersound, which refers to frequencies above 10^{13}, or 10 trillion Hz. It is almost impos-



Tiny listening devices, like the one shown here, have a host of modern applications, from electronic eavesdropping to ultrasonic stereo systems. *(Photograph by Jeffrey L. Rotman/Corbis. Reproduced by permission.)*

sible for hypersound waves to travel through most media, because its wavelengths are so short.)

What Makes the Glass Shatter

The lowest note of the eighty-eight keys on a piano is 27 Hz and the highest 4,186 Hz. This places the middle and upper register of the piano well within the optimal range for audibility, which is between 3,000 and 4,000 Hz. Clearly, the higher the note, the higher the frequency—but it is not high frequency, per se, that causes a glass to shatter when a singer or a violinist hits a certain note. All objects, or at least all rigid ones, possess their own natural frequency of vibration or oscillation. This frequency depends on a number of factors, including material composition and shape, and its characteristics are much more complex than those of sound frequency described above. In any case, a musician cannot cause a glass to shatter simply by hitting a very high note; rather, the note must be on the exact frequency at which the glass itself oscillates. Under such conditions, all the energy from the voice or musical instrument is transferred to the glass, a sudden burst that overloads the object and causes it to shatter.

To create ultrasonic waves, technicians use a transducer, a device that converts energy into ultrasonic sound waves. The most basic type of transducer is mechanical, involving oscillators or vibrating blades powered either by gas or the pressure of gas or liquids—that is, pneumatic or hydrodynamic pressure, respectively. The vibrations from these mechanical devices are on a relatively low ultrasonic frequency, and most commonly they are applied in industry for purposes such as drying or cleaning.

An electromechanical transducer, which has a much wider range of applications, converts electrical energy, in the form of current, to mechanical energy—that is, sound waves. This it does either by a magnetostrictive or a piezoelectric device. The term "magnetostrictive" comes from *magneto,* or magnetic, and *strictio,* or "drawing together." This type of transducer involves the magnetization of iron or nickel, which causes a change in dimension by forcing the atoms together. This change in dimension in turn produces a high-frequency vibration. Again, the frequency is relatively low in ultrasonic terms, and likewise the application is primarily industrial, for purposes such as cleaning and machining.

Most widely used is a transducer equipped with a specially cut piezoelectric quartz crystal. Piezoelectricity involves the application of mechanical pressure to a nonconducting crystal, which results in polarization of electrical charges, with all positive charges at one end of the crystal and all negative charges at the other end. By successively compressing and stretching the crystal at an appropriate frequency, an alternating electrical current is generated that can be converted into mechanical energy—specifically, ultrasonic waves.

Scientists use different shapes and materials (including quartz and varieties of ceramic) in fashioning piezoelectric crystals: for instance, a concave shape is best for an ultrasonic wave that will be focused on a very tight point. Piezoelectric transducers have a variety of applications in ultrasonic technology, and are capable of acting as receivers for ultrasonic vibrations.

REAL-LIFE APPLICATIONS

Pets and Pests: Ultrasonic Behavior Modification

Some of the simplest ultrasonic applications build on the fact that the upper range of audibility for human beings is relatively low among animals. Cats, by comparison, have an infrasound threshold only slightly higher than that of humans (100 Hz), but their ultrasound range of audibility is much greater—32,000 Hz instead of a mere 20,000. This explains why a cat sometimes seems to respond mysteriously to noises its owner is incapable of hearing.

For dogs, the difference is even more remarkable: their lower threshold is 40 Hz, and their high end 46,000 Hz, giving them a range more than twice that of humans. It has been said Paul McCartney, who was fond of his sheepdog Martha, arranged for the Beatles' sound engineer at Abbey Road Studios to add a short 20,000-Hz tone at the very end of *Sgt. Peppers' Lonely Hearts Club Band* in 1967. Thus—if the story is true—the Beatles' human fans would never hear the note, but it would be a special signal to Martha and all the dogs of England.

On a more practical level, a dog whistle is an extremely simple ultrasonic or near-ultrasonic device, one that obviously involves no transducers. The owner blows the whistle, which utters a tone nearly inaudible to humans but—like McCartney's 20,000-Hz tone—well within a dog's range. In fact, the Acme Silent Dog Whistle, which the company has produced since 1935, emits a tone that humans can hear (the listed range is 5,800 to 12,400 Hz), but which dogs can hear much better.

There are numerous products on the market that use ultrasonic waves for animal behavior modification of one kind or another; however, most such items are intended to repel rather than attract the animal. Hence, there are ultrasonic devices to discourage animals from relieving themselves in the wrong places, as well as some which keep unwanted dogs and cats away.

Then there is one of the most well-known uses of ultrasound for pets, which, rather than keeping other animals out, is designed to keep one's own animals in the yard. Many people know this item as an "Invisible Fence," though in fact that term is a registered trademark of the Invisible Fence Company. The "Invisible Fence" and similar products literally create a barrier of sound, using both radio signals and ultrasonics. The pet is outfitted with a collar that contains a radio receiver, and a radio transmitter is placed in some centrally located place on the owner's property—a basement, perhaps, or a garage. The "fence" itself is "visible," though usually buried, and consists of an antenna wire at the perimeter of the property. The transmitter sends a signal to the wire, which in turn signals the pet's collar. A tiny computer in the collar emits an ultrasonic sound if the animal tries to stray beyond the boundaries.

Not all animals have a higher range of hearing than humans: elephants, for instance, cannot hear tones above 12,000 Hz. On the other hand, some are drastically more sensitive acoustically than dogs: bats, whales, and dolphins all have an upper range of 150,000 Hz, though both have a low-end threshold of 1,000. Mice, at 100,000 Hz, are also at the high end, while a number of other pests—rodents and insects—fall into the region between 40,000 and 100,000 Hz. This fact has given rise to another type of ultrasonic device, for repelling all kinds of unwanted household creatures by bombarding them with ear-splitting tones.

An example of this device is the Transonic 1X-L, which offers three frequency ranges: "loud mode" (1,000-50,000 Hz); "medium mode" (10,000-50,000 Hz); and "quiet mode" (20,000-50,000 Hz). The lowest of these can be used for repelling pest birds and small animals, the medium range for insects, and the "quiet mode" for rodents.

Ultrasonic Detection in Medicine

Medicine represents one of the widest areas of application for ultrasound. Though the machinery used to provide parents-to-be with an image of their unborn child is the most well-known form of medical ultrasound, it is far from the only one. Developed in 1957 by British physician Ian Donald (1910-1987), also a pioneer in the use of ultrasonics to detect flaws in machinery, ultrasound was first used to diagnosis a patient's heart condition. Within a year, British hospitals began using it with pregnant women.

High-frequency waves penetrate soft tissue with ease, but they bounce off of harder tissue such as organs and bones, and thus send back a message to the transducer. Because each type of tissue absorbs or deflects sound differently, according to its density, the ultrasound machine can interpret these signals, creating an image of what it "sees" inside the patient's body. The technician scans the area to be studied with a series of ultrasonic waves in succession, and this results in the creation of a moving picture. It is this that creates the sight so memorable in the lives of many a modern parent: their first glimpse of their child in its mother's womb.

Though ultrasound enables physicians and nurses to determine the child's sex, this is far from being the only reason it is used. It also gives them data concerning the fetus's size; position (for instance, if the head is in a place that suggests the baby will have to be delivered by means of cesarean section); and other abnormalities.

The beauty of ultrasound is that it can provide this information without the danger posed by x rays or incisions. Doctors and ultrasound technicians use ultrasonic technology to detect body parts as small as 0.004 in (0.1 mm), making it possible to conduct procedures safely, such as locating foreign objects in the eye or measuring the depth of a severe burn. Furthermore, ultrasonic microscopes can image cellular structures to within 0.2 microns (0.002 mm).

Ultrasonic heart examination can locate tumors, valve diseases, and accumulations of fluid. Using the Doppler effect—the fact that a sound's perceived frequency changes as its source moves past the observer—physicians observe shifts in the frequency of ultrasonic measurements to determine the direction of blood flow in the body. Not only can ultrasound be used to differentiate tumors from healthy tissue, it can sometimes be used to destroy those tumors. In some cases, ultrasound actually destroys cancer cells, making use of a principle called cavitation— a promising area of ultrasound research.

Perhaps the best example of cavitation occurs when you are boiling a pot of water: bubbles—temporary cavities in the water itself—rise up from the bottom to the surface, then collapse, making a popping sound as they do. Among the research areas combining cavitation and ultrasonics are studies of light emissions produced in the collapse of a cavity created by an ultrasonic wave. These emissions are so intense that for an infinitesimal moment, they produce heat of staggering proportions—hotter than the surface of the Sun, some scientists maintain. (Again, it should be stressed that this occurs during a period too small to measure with any but the most sophisticated instruments.)

As for the use of cavitation in attacking cancer cells, ultrasonic waves can be used to create microscopic bubbles which, when they collapse, produce intense shock waves that destroy the cells. Doctors are now using a similar technique against gallstones and kidney stones. Other medical uses of ultrasound technology include ultrasonic heat for treating muscle strain, or—in a process similar to some industrial applications—the use of 25,000-Hz signals to clean teeth.

Sonar and Other Detection Devices

Airplane pilots typically use radar, but the crew of a ocean-going vessel relies on sonar (SOund Navigation and Ranging) to guide their vessel through the ocean depths. This technology takes advantage of the fact that sound waves travel well under water—much better, in fact, than light waves. Whereas a high-powered light would be of limited value underwater, particularly in the murky realms of the deep sea, sonar provides excellent data on the water's depth, as well as the location of shipwrecks, large obstacles—and, for commercial or even recreational fishermen—the presence of fish.

At the bottom of the craft's hull is a transducer, which emits an ultrasonic pulse. These sound waves travel through the water to the bottom, where they bounce back. Upon receiving the echo, the transducer sends this information to an onboard computer, which converts data on the amount of time the signal took, providing a reading of distance that gives an accurate measurement of the vessel's clearance. For instance, it takes one second for sound waves from a depth of 2,500 ft (750 m) to return to the ship. The onboard computer converts this data into a rough picture of what lies below: the ocean floor, and schools of fish or other significant objects between it and the ship.

Even more useful is a scanning sonar, which adds dimension to the scope of the ship's ultrasonic detection: not only does the sonar beam move forward along with the vessel, but it moves from side-to-side, providing a picture of a wider area along the ship's path. Sonar in general, and particularly scanning sonar, is of particular importance to a submarine's crew. Despite the fact that the periscope is perhaps the most notable feature of these underwater craft, from the viewpoint of a casual observer, in fact, the purely visual data provided by the periscope is of limited value—and that value decreases as the sub descends. It is thanks to sonar (which produces the pinging sound one so often hears in movie scenes depicting the submarine control room), combined with nuclear technology, that makes it possible for today's U.S. Navy submarines to stay submerged for months.

Sonar is perhaps the most dramatic use of ultrasonic technology for detection; less well-known—but equally intriguing, especially for its connection with clandestine activity—is the use of ultrasonics for electronic eavesdropping. Private detectives, suspicious spouses, and no doubt international spies from the CIA or Britain's MI5, use ultrasonic waves to listen to conversations in places where they cannot insert a microphone. For example, an operative might want to listen in on an encounter taking place on the seventh floor of a building with heavy security, meaning it would be impossible to plant a microphone either inside the room or on the window ledge.

Instead, the operative uses ultrasonic waves, which a transducer beams toward the window of the room being monitored. If people are speaking inside the room, this will produce vibrations on the window the transducer can detect, although the sounds would not be decipherable as conversation by a person with unaided perception. Speech vibrations from inside produce characteristic effects on the ultrasonic waves beamed back to the transducer and the operative's monitoring technology. The transducer then converts these reflected vibrations into electrical signals, which analysts can then reconstruct as intelligible sounds.

Much less dramatic, but highly significant, is the use of ultrasonic technology for detection in industry. Here the purpose is to test materials for faults, holes, cracks, or signs of corrosion. Again, the transducer beams an ultrasonic signal, and the way in which the material reflects this signal can alert the operator to issues such as metal fatigue or a faulty weld. Another method is to subject the material or materials to stress, then look for characteristic acoustic emissions from the stressed materials. (The latter is a developing field of acoustics known as acoustic emission.)

Though industrial detection applications can be used on materials such as porcelain (to test for microscopic cracks) or concrete (to evaluate how well it was poured), ultrasonics is particularly effective on metal, in which sound moves more quickly and freely than any other type of wave. Not only does ultrasonics provide an opportunity for thorough, informative, but nondestructive testing, it also allows technicians to penetrate areas where they otherwise could not go—or, in the case of ultrasonic inspection of the interior of a nuclear reactor while in operation—would not and should not go.

Binding and Loosening: A Host of Industrial Applications

Materials testing is but one among myriad uses for ultrasonics in industry, applications that can be described broadly as "binding and loosening"—either bringing materials together, or pulling them apart.

For instance, ultrasound is often used to bind, or coagulate, loose particles of dust, mist, or smoke. This makes it possible to clean a factory smokestack before it exhales pollutants into the atmosphere, or to clear clumps of fog and mist off a runway. Another form of "binding" is the use of ultrasonic vibrations to heat and weld together materials. Ultrasonics provides an even, localized flow of molten material, and is effective both on plastics and metals.

Ultrasonic soldering implements the principle of cavitation, producing microscopic bubbles in molten solder, a process that removes metal oxides. Hence, this is a case of both "binding" (soldering) and "loosening"—removing impurities from the area to be soldered. The dairy industry, too, uses ultrasonics for both purposes: ultrasonic waves break up fat globules in milk, so that the fat can be mixed together with the milk in the well-known process of homogenization. Similarly, ultrasonic pasteurization facilitates the separation of the milk from harmful bacteria and other microorganisms.

The uses of ultrasonics to "loosen" include ultrasonic humidification, wherein ultrasonic vibrations reduce water to a fine spray. Similarly, ultrasonic cleaning uses ultrasound to break down the attraction between two different types of materials. Though it is not yet practical for home use, the technology exists today to use ultrasonics for laundering clothes without using water: the ultrasonic vibrations break the bond between dirt particles and the fibers of a garment, shaking loose the dirt and subjecting the fabric to far less trauma than the agitation of a washing machine does.

As noted earlier, dentists use ultrasound for cleaning teeth, another example of loosening the bond between materials. In most of these forms of ultrasonic cleaning, a critical part of the process is the production of microscopic shock waves in the process of cavitation. The frequency of sound waves in these operations ranges from 15,000 Hz (15 kHz) to 2 million Hz (2 MHz). Ultrasonic cleaning has been used on metals, plastics, and ceramics, as well as for cleaning precision instruments used in the optical, surgical, and dental fields. Nor is it just for small objects: the electronics, automotive, and aircraft industries make heavy use of ultrasonic cleaning for a variety of machines.

Ultrasonic "loosening" makes it possible to drill though extremely hard or brittle materials, including tungsten carbide or precious stones. Just as a dental hygienist cleaning a person's teeth bombards the enamel with gentle abrasives, this form of high-intensity drilling works hand-in-hand with the use of abrasive materials such as silicon carbide or aluminum oxide.

A World of Applications

Scientists often use ultrasound in research, for instance to break up high molecular weight polymers, thus creating new plastic materials. Indeed, ultrasound also makes it possible to determine the molecular weight of liquid polymers, and to conduct other forms of investigation on the physical properties of materials.

Ultrasonics can also speed up certain chemical reactions. Hence, it has gained application in agriculture, thanks to research which revealed that seeds subjected to ultrasound may germinate more rapidly and produce higher yields. In addition to its uses in the dairy industry, noted above, ultrasonics is of value to farmers in the related beef industry, who use it to measure cows' fat layers before taking them to market.

In contrast to the use of ultrasonics for electronic eavesdropping, as noted earlier, today ultrasonic technology is available to persons who think someone might be spying on them: now they can use ultrasonics to detect the presence of electronic eavesdropping, and thus circumvent it. Closer to home is another promising application of ultrasonics for remote sensing of sounds: ultrasonic stereo speakers.

These make use of research dating back to the 1960s, which showed that ultrasound waves of relatively low frequency can carry audible sound to pinpointed locations. In 1996, Woody Norris had perfected the technology necessary to reduce distortion, and soon he and his son Joe began selling the ultrasonic speakers through the elder Norris's company, American Technology Corporation of San Diego, California.

Eric Niiler in *Business Week* described a demonstration: "Joe Norris twists a few knobs on a receiver, takes aim with a 10-inch-square gold-covered flat speaker, and blasts an invisible beam.... Thirty feet away, the tinny but easily recognizable sound of Vivaldi's *Four Seasons* rushes over you. Step to the right or left, however, and it fades away. The exotic-looking speaker emits `sound beams' that envelop the listener but are silent to those nearby. `We use the air as our virtual speakers,' says Norris...." Niiler went to note several other applications suggested by Norris: "Airline passengers could listen to their own

KEY TERMS

FREQUENCY: The number of waves passing through a given point during the interval of one second. The higher the frequency, the shorter the wavelength.

HERTZ: A unit for measuring frequency, equal to one cycle per second. If a sound wave has a frequency of 20,000 Hz, this means that 20,000 waves are passing through a given point during the interval of one second. Higher frequencies are expressed in terms of kilohertz (kHz; 103 or 1,000 cycles per second) or megahertz (MHz; 106 or 1 million cycles per second.) Hence 20,000 Hz—the threshold of ultrasonic sound—would be rendered as 20 kHz.

INFRASOUND: Sound of a frequency between 20 Hz, which places it outside the range of audibility for human beings. Its opposite is ultrasound.

LONGITUDINAL WAVE: A wave in which the individual segments vibrate in the same direction as the wave itself. This is in contrast to a transverse wave, or one in which the vibration or harmonic motion occurs perpendicular to the direction in which the wave is moving. Waves on the ocean are an example of transverse waves; by contrast, the shock waves of an explosion, the concentric waves of a radio transmission, and sound waves are all examples of longitudinal waves.

TRANSDUCER: A device that converts energy into ultrasonic sound waves.

ULTRASOUND: Sound waves with a frequency above 20,000 Hz, which makes them inaudible to the human ear. Its opposite is infrasound.

WAVELENGTH: The distance, measured on a plane parallel to that of the wave itself, between a crest and the adjacent crest, or the trough and an adjacent trough. On a longitudinal wave, this is simply the distance between waves, which constitute a series of concentric circles radiating from the source.

WAVE MOTION: Activity that carries energy from one place to another without actually moving any matter.

music channel*sans* headphones without disturbing neighbors. Troops could confuse the enemy with `virtual' artillery fire, or talk to each other without having their radio communications picked up by eavesdroppers."

WHERE TO LEARN MORE

Beiser, Arthur. *Physics*, 5th ed. Reading, MA: Addison-Wesley, 1991.

Crocker, Malcolm J. *Encyclopedia of Acoustics*. New York: John Wiley & Sons, 1997.

Knight, David C. *Silent Sound: The World of Ultrasonics*. New York: Morrow, 1980.

Langone, John. *National Geographic's How Things Work*. Washington, D.C.: National Geographic Society, 1999.

Medical Ultrasound WWW Directory (Web site). <http://www.ultrasoundinsider.com> (February 16, 2001).

Niiler, Eric; edited by Alex Salkever. "Now Here This—If You're in the Sweet Spot." *Business Week*, October 16, 2000.

Meire, Hylton B. and Pat Farrant. *Basic Ultrasound*. Chichester, N.Y.: John Wiley & Sons, 1995.

Suplee, Curt. *Everyday Science Explained*. Washington, D.C.: National Geographic Society, 1996.

LIGHT AND ELECTROMAGNETISM

MAGNETISM

CONCEPT

Most people are familiar with magnets primarily as toys, or as simple objects for keeping papers attached to a metal surface such as a refrigerator door. In fact the areas of application for magnetism are much broader, and range from security to health care to communication, transportation, and numerous other aspects of daily life. Closely related to electricity, magnetism results from specific forms of alignment on the part of electron charges in certain varieties of metal and alloy.

HOW IT WORKS

Magnetism, along with electricity, belongs to a larger phenomenon, electromagnetism, or the force generated by the passage of an electric current through matter. When two electric charges are at rest, it appears to the observer that the force between them is merely electric. If the charges are in motion, however—and in this instance motion or rest is understood in relation to the observer—then it appears as though a different sort of force, known as magnetism, exists between them.

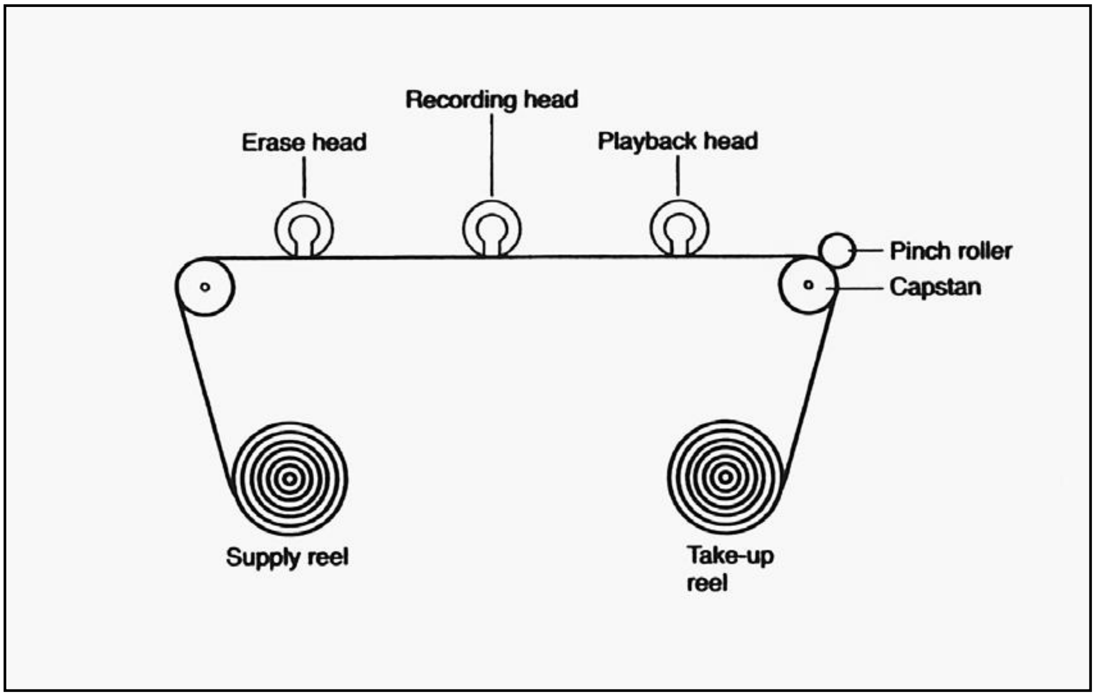
In fact, the difference between magnetism and electricity is purely artificial. Both are manifestations of a single fundamental force, with "magnetism" simply being an abstraction that people use for the changes in electromagnetic force created by the motion of electric charges. It is a distinction on the order of that between water and wetness; nonetheless, it is often useful and convenient to discuss the two phenomena as though they were separate.

At the atomic level, magnetism is the result of motion by electrons, negatively charged subatomic particles, relative to one another. Rather like planets in a solar system, electrons both revolve around the atom's nucleus and rotate on their own axes. (In fact the exact nature of their movement is much more complex, but this analogy is accurate enough for the present purposes.) Both types of movement create a magnetic force field between electrons, and as a result the electron takes on the properties of a tiny bar magnet with a north pole and south pole. Surrounding this infinitesimal magnet are lines of magnetic force, which begin at the north pole and curve outward, describing an ellipse as they return to the south pole.

In most atomic elements, the structure of the atom is such that the electrons align in a random manner, rather like a bunch of basketballs bumping into one another as they float in a swimming pool. Because of this random alignment, the small magnetic fields cancel out one another. Two such self-canceling particles are referred to as paired electrons, and again, the analogy to bar magnets is an appropriate one: if one were to shake a bag containing an even number of bar magnets, they would all wind up in pairs, joined at opposing (north-south) poles.

There are, however, a very few elements in which the fields line up to create what is known as a net magnetic dipole, or a unity of direction—rather like a bunch of basketballs simultaneously thrown from in the same direction at the same time. These elements, among them iron, cobalt, and nickel, as well as various alloys or mixtures, are commonly known as magnetic metals or natural magnets.

It should be noted that in magnetic metals, magnetism comes purely from the alignment of



THE RECORDING HEAD IS A SMALL ELECTROMAGNET WHOSE MAGNETIC FIELD EXTENDS OVER THE SECTION OF TAPE BEING RECORDED. LOUD SOUNDS PRODUCE STRONG MAGNETIC FIELDS, AND SOFT ONES WEAK FIELDS. *(Illustration by Hans & Cassidy. The Gale Group.)*

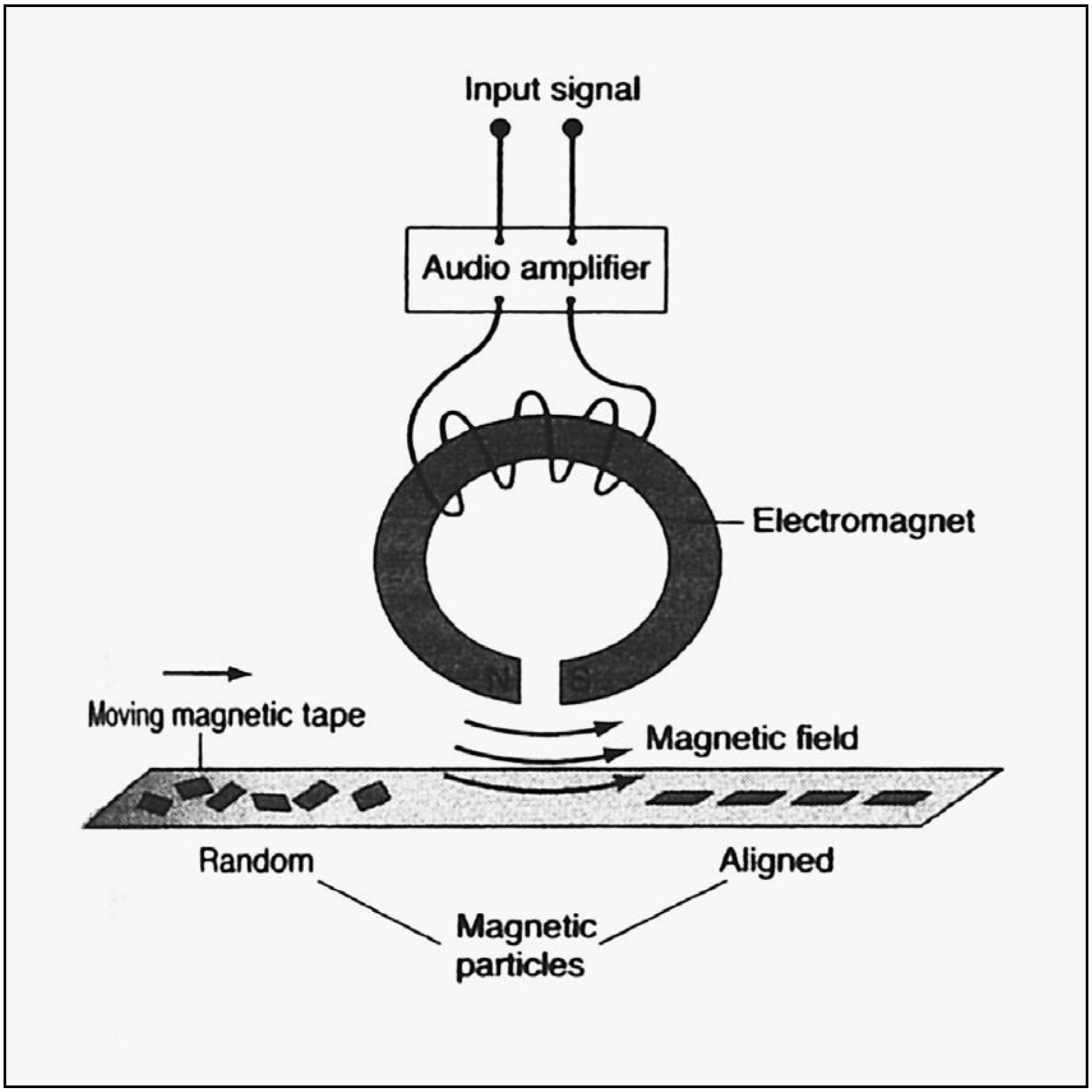
forces exerted by electrons as they spin on their axes, whereas the forces created by their orbital motion around the nucleus tend to cancel one another out. But in magnetic rare earth elements such as cerium, magnetism comes both from rotational and orbital forms of motion. Of principal concern in this discussion, however, is the behavior of natural magnets on the one hand, and of nonmagnetic materials on the other.

There are five different types of magnetism—diamagnetism, paramagnetism, ferromagnetism, ferrimagnetism, and antiferromagnetism. Actually, these terms describe five different types of response to the process of magnetization, which occurs when an object is placed in a magnetic field.

A magnetic field is an area in which a magnetic force acts on a moving charged particle such that the particle would experience no force if it moved in the direction of the magnetic field—in other words, it would be "drawn," as a ten-penny nail is drawn to a common bar or horseshoe (U-shaped) magnet. An electric current is an example of a moving charge, and indeed one of the best ways to create a magnetic field is with a current. Often this is done by means of a solenoid, a current-carrying wire coil through which the material to be magnetized is passed, much as one would pass an object through the interior of a spring.

All materials respond to a magnetic field; they just respond in different ways. Some nonmagnetic substances, when placed within a magnetic field, slightly reduce the strength of that field, a phenomenon known as diamagnetism. On the other hand, there are nonmagnetic substances possessing an uneven number of electrons per atom, and in these instances a slight increase in magnetism, known as paramagnetism, occurs. Paramagnetism always has to overcome diamagnetism, however, and hence the gain in magnetic force is very small. In addition, the thermal motion of atoms and molecules prevents the objects' magnetic fields from coming into alignment with the external field. Lower temperatures, on the other hand, enhance the process of paramagnetism.

In contrast to diamagnetism and paramagnetism, ferro-, ferri-, and antiferromagnetism all describe the behavior of natural magnets when exposed to a magnetic field. The name ferromagnetism suggests a connection with iron, but in fact the term can apply to any of those materials in which the magnitude of the object's magnetic



The permanent magnetization of a natural magnet is difficult to reverse, but reversal of a tape's magnetization—in other words, erasing the tape—is easy. An erase head, an electromagnet operating at a frequency too high for the human ear to hear, simply scrambles the magnetic particles on a piece of tape. *(Illustration by Hans & Cassidy. The Gale Group.)*

field increases greatly when it is placed within an external field. When a natural magnet becomes magnetized (that is, when a metal or alloy comes into contact with an external magnetic field), a change occurs at the level of the domain, a group of atoms equal in size to about 5×10^{-5} meters across—just large enough to be visible under a microscope.

In an unmagnetized sample, there may be an alignment of unpaired electron spins within a domain, but the direction of the various domains' magnetic forces in relation to one another is random. Once a natural magnet is placed within an external magnetic force field, however, one of two things happens to the domains. Either they all come into alignment with the field or, in certain types of material, those domains in alignment with the field grow while the others shrink to nonexistence.

The first of these processes is called domain alignment or ferromagnetism, the second domain growth or ferrimagnetism. Both processes turn a natural magnet into what is known as a permanent magnet—or, in common parlance, simply a "magnet." The latter is then capable of temporarily magnetizing a ferromagnetic item, as for instance when one rubs a paper clip against a permanent magnet and then uses the magnetized clip to lift other paper clips. Of the two varieties, however, a ferromagnetic metal is stronger because it requires a more powerful magnetic force field in order to become magnetized. Most powerful of all is a saturated ferromagnetic metal, one in which all the unpaired electron spins are aligned.

Once magnetized, it is very hard for a ferromagnetic metal to experience demagnetization, or antiferromagnetism. Again, there is a connection between temperature and magnetism, with heat acting as a force to reduce the strength of a magnetic field. Thus at temperatures above 1,418°F (770°C), the atoms within a domain take on enough kinetic energy to overpower the forces holding the electron spins in alignment. In addition, mechanical disturbances—for instance, battering a permanent magnet with a hammer—can result in some reduction of magnetic force.

Many of the best permanent magnets are made of steel, which, because it is an alloy of iron with carbon and other elements, has an irregular structure that lends itself well to the ferromagnetic process of domain alignment. Iron, by contrast, will typically lose its magnetization when an external magnetic force field is removed; but this actually makes it a better material for some varieties of electromagnet.

The latter, in its simplest form, consists of an iron rod inside a solenoid. When a current is passing through the solenoid, it creates a magnetic force field, activating the iron rod and turning it into an electromagnet. But as soon as the current is turned off, the rod loses its magnetic force. Not only can an electromagnet thus be controlled, but it is often stronger than a permanent magnet: hence, for instance, giant electromagnets are used for lifting cars in junkyards.

REAL-LIFE APPLICATIONS

Finding the Way: Magnets in Compasses

A north-south bar magnet exerts exactly the same sort of magnetic field as a solenoid. Lines of magnetic run through it in one direction, from "south" to "north," and upon leaving the north pole of the magnet, these lines describe an ellipse as they curve back around to the south pole. In view of this model, it is also easy to comprehend why a pair of opposing poles attracts one another, and a pair of like poles—for whom the lines of force are moving away from each other—repels. This is a fact particularly applicable to the operation of MAGLEV trains, as discussed later.

A magnetic compass works because Earth itself is like a giant bar magnet, complete with vast arcs of magnetic force, called the geomagnetic field, surrounding the planet. The first scientist to recognize the magnetic properties of Earth was the English physicist William Gilbert (1544-1603). Scientists today believe that the source of Earth's magnetism lies in a core of molten iron some 4,320 mi (6,940 km) across, constituting half the planet's diameter. Within this core run powerful electric currents that ultimately create the geomagnetic field.

Just as a powerful magnet causes all the domains in a magnetic metal to align with it, a bar magnet placed in a magnetic field will rotate until it lines up with the field's direction. The same thing happens when one suspends a magnet from a string: it lines up with Earth's magnetic field, and points in a north-south direction. The Chinese of the first century B.C., though unaware of the electromagnetic forces that caused this to happen, discovered that a strip of magnetic metal always tended to point toward geographic north.

This led ultimately to the development of the magnetic compass, which typically consists of a magnetized iron needle suspended over a card marked with the four cardinal directions. The needle is attached to a pivoting mechanism at its center, which allows it to move freely so that the "north" end will always point the user northward.

The magnetic compass proved so important that it is typically ranked alongside paper, printing, and gunpowder as one of premodern China's four great gifts to the West. Prior to the compass, mariners had to depend purely on the position of the Sun and other, less reliable, means of determining direction; hence the invention quite literally helped open up the world. But there is a somewhat irksome anomaly lurking in the seeming simplicity of the magnetic compass.

In fact magnetic north is not the same thing as true north; or, to put it another way, if one continued to follow a compass northward, it would lead not to the Earth's North Pole, but to a point identified in 1984 as 77°N, 102°18' W—that is, in the Queen Elizabeth Islands of far northern Canada. The reason for this is that Earth's magnetic field describes a current loop whose center is 11° off the planet's equator, and thus the north and south magnetic poles—which are on a plane perpendicular to that of the Earth's magnetic field—are 11° off of the planet's axis.

The magnetic field of Earth is changing position slowly, and every few years the United States Geological Survey updates magnetic declination, or the shift in the magnetic field. In addition, Earth's magnetic field is slowly weakening as well. The behavior, both in terms of weakening and movement, appears to be similar to changes taking place in the magnetic field of the Sun.

Magnets for Detection: Burglar Alarms, Magnetometers, and MRI

A compass is a simple magnetic instrument, and a burglar alarm is not much more complex. A magnetometer, on the other hand, is a much more sophisticated piece of machinery for detecting the strength of magnetic fields. Nonetheless, the magnetometer bears a relation to its simpler cousins: like a compass, certain kinds of magnetometers respond to a planet's magnetic field; and like a burglar alarm, other varieties of magnetometer are employed for security.

At heart, a burglar alarm consists of a contact switch, which responds to changes in the environment and sends a signal to a noisemaking device. The contact switch may be mechanical—a simple fastener, for instance—or magnetic. In the latter case, a permanent magnet may be installed in the frame of a window or door, and a piece of magnetized material in the window or door itself. Once the alarm is activated, it will respond to any change in the magnetic field—i.e., when someone slides open the door or window, thus breaking the connection between magnet and metal.

Though burglar alarms may vary in complexity, and indeed there may be much more advanced systems using microwaves or infrared rays, the application of magnetism in home security is a simple matter of responding to changes in a magnetic field. In this regard, the principle governing magnetometers used at security checkpoints is even simpler. Whether at an airport or at the entrance to some other high-security venue, whether handheld or stationary, a magnetometer merely detects the presence of magnetic metals. Since the vast majority of firearms, knife-blades, and other weapons are made of iron or steel, this provides a fairly efficient means of detection.

At a much larger scale, magnetometers used by astronomers detect the strength and sometimes the direction of magnetic fields surrounding Earth and other bodies in space. This variety of magnetometer dates back to 1832, when mathematician and scientist Carl Friedrich Gauss (1777-1855) developed a simple instrument consisting of a permanent bar magnet suspended horizontally by means of a gold wire. By measuring the period of the magnet's oscillation in Earth's magnetic field (or magnetosphere), Gauss was able to measure the strength of that field. Gauss's name, incidentally, would later be applied to the term for a unit of magnetic force. The gauss, however, has in recent years been largely replaced by the tesla, named after Nikola Tesla (1856-1943), which is equal to one newton/ampere meter (1 N/A•m) or 10^4 (10,000) gauss.

As for magnetometers used in astronomical research, perhaps the most prominent—and certainly one of the most distant—ones is on *Galileo*, a craft launched by the U.S. National Aeronautics and Space Administration (NASA) toward Jupiter on October 15, 1989. Among other instruments on board *Galileo*, which has been in orbit around the solar system's largest planet since 1995, is a magnetometer for measuring Jupiter's magnetosphere and that of its surrounding asteroids and moons.

Closer to home, but no less impressive, is another application of magnetism for the purposes of detection: magnetic resonance imagining, or MRI. First developed in the early 1970s, MRI permits doctors to make intensive diagnoses without invading the patient's body either with a surgical knife or x rays.

The heart of the MRI machine is a large tube into which the patient is placed in a supine position. A technician then activates a powerful magnetic field, which causes atoms within the patient's body to spin at precise frequencies. The machine then beams radio signals at a frequency matching that of the atoms in the cells (e.g., cancer cells) being sought. Upon shutting off the radio signals and magnetic field, those atoms emit bursts of energy that they have absorbed from the radio waves. At that point a computer scans the body for frequencies matching specific types of atoms, and translates these into three-dimensional images for diagnosis.

Magnets for Projecting Sound: Microphones, Loudspeakers, Car Horns, and Electric Bells

The magnets used in *Galileo* or an MRI machine are, needless to say, very powerful ones, and as noted earlier, the best way to create a super-strong, controllable magnet is with an electrical current. When that current is properly coiled around a magnetic metal, this creates an electromagnet, which can be used in a variety of applications.

As discussed above, the most powerful electromagnets typically use nonpermanent magnets so as to facilitate an easy transition from an extremely strong magnetic field to a weak or nonexistent one. On the other hand, permanent magnets are also used in loudspeakers and similar electromagnetic devices, which seldom require enormous levels of power.

In discussing the operation of a loudspeaker, it is first necessary to gain a basic understanding of how a microphone works. The latter contains a capacitor, a system for storing charges in the form of an electrical field. The capacitor's negatively charged plate constitutes the microphone's diaphragm, which, when it is hit by sound waves, vibrates at the same frequency as those waves. Current flows back and forth between the diaphragm and the positive plate of the capacitor, depending on whether the electrostatic or electrical pull is increasing or decreasing. This in turn produces an alternating current, at the same frequency as the sound waves, which travels through a mixer and then an amplifier to the speaker.

A loudspeaker typically contains a circular permanent magnet, which surrounds an electrical coil and is in turn attached to a cone-shaped diaphragm. Current enters the speaker ultimately from the microphone, alternating at the same frequency as the source of the sound (a singer's voice, for instance). As it enters the coil, this current induces an alternating magnetic field, which causes the coil to vibrate. This in turn vibrates the cone-shaped diaphragm, and the latter reproduces sounds generated at the source.

A car horn also uses magnetism to create sound by means of vibration. When a person presses down on the horn embedded in his or her steering wheel, this in turn depresses an iron bar that passes through an electromagnet surrounded by wires from the car's battery. The bar moves up and down within the electromagnetic field, causing the diaphragm to vibrate and producing a sound that is magnified greatly when released through a bell-shaped horn.

Electromagnetically induced vibration is also the secret behind another noise-making device, a vibrating electric doorbell used in many apartments. The button that a visitor presses is connected directly to a power source, which sends current flowing through a spring surrounding an electromagnet. The latter generates a magnetic field, drawing toward it an iron armature attached to a hammer. The hammer then strikes the bell. The result is a mechanical reaction that pushes the armature away from the electromagnet, but the spring forces the armature back against the electromagnet again. This cycle of contact and release continues for as long as the button is depressed, causing a continual ringing of the bell.

Recording and Reading Data Using Magnets: From Records and Tapes to Disk Drives

Just as magnetism plays a critical role in projecting the volume of sound, it is also crucial to the recording and retrieval of sound and other data. Of course terms such as "retrieval" and "data" have an information-age sound to them, but the idea of using magnetism to record sound is an old one—much older than computers or compact discs (CDs). The latter, of course, replaced cassettes in the late 1980s as the preferred mode for listening to recorded music, just as cassettes had recently made powerful gains against phonograph records.

Despite the fact that cassettes entered the market much later than records, however, recording engineers from the mid-twentieth century onward typically used magnetic tape for master recordings of songs. This master would then be used to create a metal master record disk by means of a cutting head that responded to vibrations from the master tape; then, the record company could produce endless plastic copies of the metal record.

In recording a tape—whether a stereo master or a mere home recording of a conversation—the principles at work are more or less the same. As noted in the earlier illustration involving a microphone and loudspeaker, sound comes through a microphone in the form of alternating current. The strength of this current in turn affects the "recording head," a small electromagnet whose magnetic field extends over the section of tape being recorded. Loud sounds produce strong magnetic fields, and soft ones weak fields.

All of this information becomes embedded on the cassette tape through a process of magnetic alignment not so different from the process described earlier for creating a permanent magnet. But whereas the permanent magnetization of a natural magnet is difficult to reverse, reversal of a tape's magnetization—in other words, erasing the tape—is easy. An erase head, an electromagnet operating at a frequency too high for the human ear to hear, simply scrambles the magnetic particles on a piece of tape.

A CD, as one might expect, is much more closely related to a computer disk-drive than it is to earlier forms of recording technology. The disk drive receives electronic on-off signals from the computer, and translates these into magnetic codes that it records on the surface of a floppy disk. The disk drive itself includes two electric motors: a disk motor, which rotates the disk at a high speed, and a head motor, which moves the computer's read-write head across the disk. (It should be noted that most electric motors, including the universal motors used in a variety of household appliances, also use electromagnets.)

A third motor, called a stepper motor, ensures that the drive turns at a precise rate of speed. The stepper motor contains its own magnet, in this case a permanent one of cylindrical shape that sends signals to rows of metal teeth surrounding it, and these teeth act as gears to regulate the drive's speed. Likewise a CD player, which actually uses laser beams rather than magnetic fields to retrieve data from a disc, also has a drive system that regulates the speed at which the disk spins.

MAGLEV TRAINS, LIKE THIS EXPERIMENTAL ONE IN MIYAZAKI, JAPAN, MAY REPRESENT THE FUTURE OF MASS TRANSIT. *(Photograph by Michael S. Yamashita/Corbis. Reproduced by permission.)*

MAGLEV Trains: The Future of Transport?

One promising application of electromagnetic technology relates to a form of transportation that might, at first glance, appear to be old news: trains. But MAGLEV, or magnetic levitation, trains are as far removed from the old steam engines of the Union Pacific as the space shuttle is from the Wright brothers' experimental airplane.

As discussed earlier, magnetic poles of like direction (i.e., north-north or south-south) repel one another such that, theoretically at least, it is possible to keep one magnet suspended in the air over another magnet. Actually it is impossible to produce these results with simple bar magnets, because their magnetic force is too small; but an electromagnet can create a magnetic field powerful enough that, if used properly, it exerts enough repulsive force to lift extremely heavy objects. Specifically, if one could activate train tracks with a strong electromagnetic field, it might be possible to "levitate" an entire train. This in turn

KEY TERMS

ELECTROMAGNET: A type of magnet in which an object is charged by an electrical current. Typically the object used is made of iron, which quickly loses magnetic force when current is reduced. Thus an electromagnet can be turned on or off, and its magnetic force altered, making it potentially much more powerful than a natural magnet.

ELECTROMAGNETISM: The unified electrical and magnetic force field generated by the passage of an electric current through matter.

ELECTRONS: Negatively charged subatomic particles whose motion relative to one another creates magnetic force.

MAGNETIC FIELD: Wherever a magnetic force acts on a moving charged particle, a magnetic field is said to exist. Magnetic fields are typically measured by a unit called a tesla.

NATURAL MAGNET: A chemical element in which the magnetic fields created by electrons' relative motion align uniformly to create a net magnetic dipole, or unity of direction. Such elements, among them iron, cobalt, and nickel, are also known as magnetic metals.

PERMANENT MAGNET: A magnetic material in which groups of atoms, known as domains, are brought into alignment, and in which magnetization cannot be changed merely by attempting to realign the domains. Permanent magnetization is reversible only at very high temperatures—for example, 1,418°F (770°C) in the case of iron.

would make possible a form of transport that could move large numbers of people in relative comfort, thus decreasing the environmental impact of automobiles, and do so at much higher speeds than a car could safely attain.

Actually the idea of MAGLEV trains goes back to a time when trains held complete supremacy over automobiles as a mode of transportation: specifically, 1907, when rocket pioneer Robert Goddard (1882-1945) wrote a story describing a vehicle that traveled by means of magnetic levitation. Just five years later, French engineer Emile Bachelet produced a working model for a MAGLEV train. But the amount of magnetic force required to lift such a vehicle made it impractical, and the idea fell to the wayside.

Then, in the 1960s, the advent of superconductivity—the use of extremely low temperatures, which facilitate the transfer of electrical current through a conducting material with virtually no resistance—made possible electromagnets of staggering force. Researchers began building MAGLEV prototypes using superconducting coils with strong currents to create a powerful magnetic field. The field in turn created a repulsive force capable of lifting a train several inches above a railroad track. Electrical current sent through guideway coils on the track allowed for enormous propulsive force, pushing trains forward at speeds up to and beyond 250 MPH (402 km/h).

Initially, researchers in the United States were optimistic about MAGLEV trains, but safety concerns led to the shelving of the idea for several decades. Meanwhile, other industrialized nations moved forward with MAGLEVs: in Japan, engineers built a 27-mi (43.5-km) experimental MAGLEV line, while German designers experimented with attractive (as opposed to repulsive) force in their Transrapid 07. MAGLEV trains gained a new defender in the United States with now-retired Senator Daniel Patrick Moynihan (D-NY), who as chairman of a Senate subcommittee overseeing the interstate highway system introduced legislation to fund MAGLEV research. The 1998 transportation bill allocated $950 million toward the Magnetic Levitation Prototype Development Program. As part of this program, in January 2001 the U.S. Department of Transportation selected projects in Maryland and Pennsylvania as the two finalists in the competition to build the first MAGLEV train service in the United States. The goal is to have the service in place by approximately 2010.

WHERE TO LEARN MORE

Barr, George. *Science Projects for Young People.* New York: Dover, 1964.

Beiser, Arthur. *Physics,* 5th ed. Reading, MA: Addison-Wesley, 1991.

Hann, Judith. *How Science Works.* Pleasantville, NY: Reader's Digest, 1991.

Macaulay, David. *The New Way Things Work.* Boston: Houghton Mifflin, 1998.

Molecular Expressions: Electricity and Magnetism: Interactive Java Tutorials (Web site). <http://micro.magnet.fsu.edu/electromag/java/> (January 26, 2001).

Topical Group on Magnetism (Web site). <http://www.aps.org/units/gmag/> (January 26, 2001).

VanCleave, Janice. *Magnets.* New York: John Wiley & Sons, 1993.

Wood, Robert W. *Physics for Kids: 49 Easy Experiments with Electricity and Magnetism.* New York: Tab, 1990.

ELECTROMAGNETIC SPECTRUM

CONCEPT

One of the most amazing aspects of physics is the electromagnetic spectrum—radio waves, microwaves, infrared light, visible light, ultraviolet light, x rays, and gamma rays—as well as the relationship between the spectrum and electromagnetic force. The applications of the electromagnetic spectrum in daily life begin the moment a person wakes up in the morning and "sees the light." Yet visible light, the only familiar part of the spectrum prior to the eighteenth and nineteenth centuries, is also its narrowest region. Since the beginning of the twentieth century, uses for other bands in the electromagnetic spectrum have proliferated. At the low-frequency end are radio, short-wave radio, and television signals, as well as the microwaves used in cooking. Higher-frequency waves, all of which can be generally described as light, provide the means for looking deep into the universe—and deep into the human body.

HOW IT WORKS

Electromagnetism

The ancient Romans observed that a brushed comb would attract particles, a phenomenon now known as static electricity and studied within the realm of electrostatics in physics. Yet, the Roman understanding of electricity did not extend any further, and as progress was made in the science of physics—after a period of more than a thousand years, during which scientific learning in Europe progressed very slowly—it developed in areas that had nothing to do with the strange force observed by the Romans.

The fathers of physics as a serious science, Galileo Galilei (1564-1642) and Sir Isaac Newton (1642-1727), were concerned with gravitation, which Newton identified as a fundamental force in the universe. For nearly two centuries, physicists continued to believe that there was only one type of force. Yet, as scientists became increasingly aware of molecules and atoms, anomalies began to arise—in particular, the fact that gravitation alone could not account for the strong forces holding atoms and molecules together to form matter.

FOUNDATIONS OF ELECTROMAGNETIC THEORY. At the same time, a number of thinkers conducted experiments concerning the nature of electricity and magnetism, and the relationship between them. Among these were several giants in physics and other disciplines—including one of America's greatest founding fathers. In addition to his famous (and highly dangerous) experiment with lightning, Benjamin Franklin (1706-1790) also contributed the names "positive" and "negative" to the differing electrical charges discovered earlier by French physicist Charles Du Fay (1698-1739).

In 1785, French physicist and inventor Charles Coulomb (1736-1806) established the basic laws of electrostatics and magnetism. He maintained that there is an attractive force that, like gravitation, can be explained in terms of the inverse of the square of the distance between objects. That attraction itself, however, resulted not from gravity, but from electrical charge, according to Coulomb.

A few years later, German mathematician Johann Karl Friedrich Gauss (1777-1855) developed a mathematical theory for finding the magnetic potential of any point on Earth, and his contemporary, Danish physicist Hans Christian Oersted (1777-1851), became the first scientist to establish the existence of a clear relationship between electricity and magnetism. This led to the foundation of electromagnetism, the branch of physics devoted to the study of electrical and magnetic phenomena.

French mathematician and physicist André Marie Ampère (1775-1836) concluded that magnetism is the result of electricity in motion, and, in 1831, British physicist and chemist Michael Faraday (1791-1867) published his theory of electromagnetic induction. This theory shows how an electrical current in one coil can set up a current in another through the development of a magnetic field. This enabled Faraday to develop the first generator, and for the first time in history, humans were able to convert mechanical energy systematically into electrical energy.

MAXWELL AND ELECTROMAGNETIC FORCE. A number of other figures contributed along the way; but, as yet, no one had developed a "unified theory" explaining the relationship between electricity and magnetism. Then, in 1865, Scottish physicist James Clerk Maxwell (1831-1879) published a groundbreaking paper, "On Faraday's Lines of Force," in which he outlined a theory of electromagnetic force—the total force on an electrically charged particle, which is a combination of forces due to electrical and/or magnetic fields around the particle.

Maxwell had thus discovered a type of force in addition to gravity, and this reflected a "new" type of fundamental interaction, or a basic mode by which particles interact in nature. Newton had identified the first, gravitational interaction, and in the twentieth century, two other forms of fundamental interaction—strong nuclear and weak nuclear—were identified as well.

In his work, Maxwell drew on the studies conducted by his predecessors, but added a new statement: that electrical charge is conserved. This statement, which did not contradict any of the experimental work done by the other physicists, was based on Maxwell's predictions regarding what should happen in situations of electromagnetism; subsequent studies have supported his predictions.

Electromagnetic Radiation

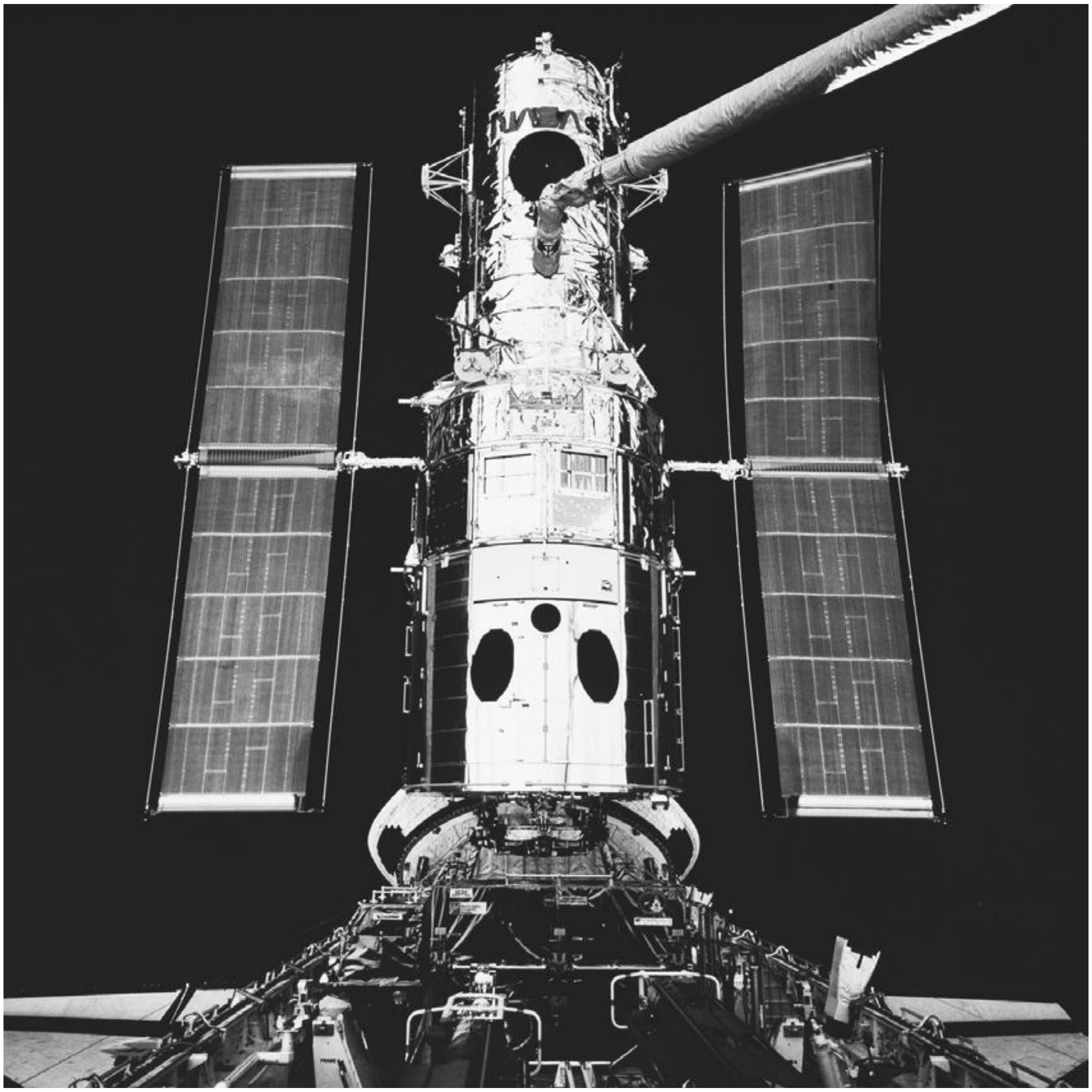
So far, what we have seen is the foundation for modern understanding of electricity and magnetism. This understanding grew enormously in the late nineteenth and early twentieth centuries, thanks both to the theoretical work of physicists, and the practical labors of inventors such as Thomas Alva Edison (1847-1931) and Serbian-American electrical engineer Nikola Tesla (1856-1943). But our concern in the present context is with electromagnetic radiation, of which the waves on the electromagnetic spectrum are a particularly significant example.

Energy can travel by conduction or convection, two principal means of heat transfer. But the energy Earth receives from the Sun—the energy conveyed through the electromagnetic spectrum—is transferred by another method, radiation. Whereas conduction of convection can only take place where there is matter, which provides a medium for the energy transfer, radiation requires no medium. Thus, electromagnetic energy passes from the Sun to Earth through the vacuum of empty space.

ELECTROMAGNETIC WAVES. The connection between electromagnetic radiation and electromagnetic force is far from obvious. Even today, few people not scientifically trained understand that there is a clear relationship between electricity and magnetism—let alone a connection between these and visible light. The breakthrough in establishing that connection can be attributed both to Maxwell and to German physicist Heinrich Rudolf Hertz (1857-1894).

Maxwell had suggested that electromagnetic force carried with it a certain wave phenomenon, and predicted that these waves traveled at a certain speed. In his *Treatise on Electricity and Magnetism* (1873), he predicted that the speed of these waves was the same as that of light—186,000 mi (299,339 km) per second—and theorized that the electromagnetic interaction included not only electricity and magnetism, but light as well. A few years later, while studying the behavior of electrical currents, Hertz confirmed Maxwell's proposition regarding the wave phenomenon by showing that an electrical current generated some sort of electromagnetic radiation.

In addition, Hertz found that the flow of electrical charges could be affected by light under certain conditions. Ultraviolet light had already been identified, and Hertz shone an ultraviolet beam on the negatively charged side of a gap in a



The Hubble Space Telescope includes an ultraviolet light instrument called the Goddard High Resolution Spectrograph that it is capable of observing extremely distant objects. *(Photograph by Roger Ressmeyer/Corbis. Reproduced by permission.)*

current loop. This made it easier for an electrical spark to jump the gap. Hertz could not explain this phenomenon, which came to be known as the photoelectric effect. Indeed, no one else could explain it until quantum theory was developed in the early twentieth century. In the meantime, however, Hertz's discovery of electromagnetic waves radiating from a current loop led to the invention of radio by Italian physicist and engineer Guglielmo Marconi (1874-1937) and others.

Light: Waves or Particles?

At this point, it is necessary to jump backward in history, to explain the progression of scientists' understanding of light. Advancement in this area took place over a long period of time: at the end of the first millennium A.D., the Arab physicist Alhasen (Ibn al-Haytham; c. 965-1039) showed that light comes from the Sun and other self-illuminated bodies—not, as had been believed up to that time—from the eye itself. Thus, studies in optics, or the study of light and vision, were—compared to understanding of electromagnetism itself—relatively advanced by 1666, when Newton discovered the spectrum of colors in light. As Newton showed, colors are arranged in a sequence, and white light is a combination of all colors.

Newton put forth the corpuscular theory of light—that is, the idea that light is made up of particles—but his contemporary Christiaan Huygens (1629-1695), a Dutch physicist and astronomer, maintained that light appears in the form of a wave. For the next century, adherents of Newton's corpuscular theory and of Huygens's wave theory continued to disagree. Physicists on the European continent began increasingly to accept wave theory, but corpuscular theory remained strong in Newton's homeland.

Thus, it was ironic that the physicist whose work struck the most forceful blow against corpuscular theory was himself an Englishman: Thomas Young (1773-1829), who in 1801 demonstrated interference in light. Young directed a beam of light through two closely spaced pinholes onto a screen, reasoning that if light truly were made of particles, the beams would project two distinct points onto the screen. Instead, what he saw was a pattern of interference—a wave phenomenon.

By the time of Hertz, wave theory had become dominant; but the photoelectric effect also exhibited aspects of particle behavior. Thus, for the first time in more than a century, particle theory gained support again. Yet, it was clear that light had certain wave characteristics, and this raised the question—which is it, a wave or a set of particles streaming through space?

The work of German physicist Max Planck (1858-1947), father of quantum theory, and of Albert Einstein (1879-1955), helped resolve this apparent contradiction. Using Planck's quantum principles, Einstein, in 1905, showed that light appears in "bundles" of energy, which travel as waves but behave as particles in certain situations. Eighteen years later, American physicist Arthur Holly Compton (1892-1962) showed that, depending on the way it is tested, light appears as either a particle or a wave. These particles he called photons.

WAVE MOTION AND ELECTROMAGNETIC WAVES

The particle behavior of electromagnetic energy is beyond the scope of the present discussion, though aspects of it are discussed elsewhere. For the present purposes, it is necessary only to view the electromagnetic spectrum as a series of waves, and in the paragraphs that follow, the rudiments of wave motion will be presented in short form.

A type of harmonic motion that carries energy from one place to another without actually moving any matter, wave motion is related to oscillation, harmonic—and typically periodic—motion in one or more dimensions. Oscillation involves no net movement, but only movement in place; yet individual waves themselves are oscillating, even as the overall wave pattern moves.

The term periodic motion, or movement repeated at regular intervals called periods, describes the behavior of periodic waves: waves in which a uniform series of crests and troughs follow each other in regular succession. Periodic waves are divided into longitudinal and transverse waves, the latter (of which light waves are an example) being waves in which the vibration or motion is perpendicular to the direction in which the wave is moving. Unlike longitudinal waves, such as those that carry sound energy, transverse waves are fairly easy to visualize, and assume the shape that most people imagine when they think of waves: a regular up-and-down pattern, called "sinusoidal" in mathematical terms.

PARAMETERS OF WAVE MOTION. A period (represented by the symbol T) is the amount of time required to complete one full cycle of the wave, from trough to crest and back to trough. Period is mathematically related to several other aspects of wave motion, including wave speed, frequency, and wavelength.

Frequency (abbreviated f) is the number of waves passing through a given point during the interval of one second. It is measured in Hertz (Hz), named after Hertz himself: a single Hertz (the term is both singular and plural) is equal to one cycle of oscillation per second. Higher frequencies are expressed in terms of kilohertz (kHz; 10^3 or 1,000 cycles per second); megahertz (MHz; 10^6 or 1 million cycles per second); and gigahertz (GHz; 10^9 or 1 billion cycles per second.)

Wavelength (represented by the symbol λ, the Greek letter lambda) is the distance between a crest and the adjacent crest, or a trough and an adjacent trough, of a wave. The higher the frequency, the shorter the wavelength; and, thus, it is possible to describe waves in terms of either. According to quantum theory, however, electromagnetic waves can also be described in terms of photon energy level, or the amount of energy in each photon. Thus, the electromagnetic spectrum, as we shall see, varies from relatively long-wavelength, low-frequency, low-energy radio waves on the one end to extremely short-wavelength, high-frequency, high-energy gamma rays on the other.

The other significant parameter for describing a wave—one mathematically independent from those so far discussed—is amplitude. Defined as the maximum displacement of a vibrating material, amplitude is the "size" of a wave. The greater the amplitude, the greater the energy the wave contains: amplitude indicates intensity. The amplitude of a light wave, for instance, determines the intensity of the light.

A RIGHT-HAND RULE. Physics textbooks use a number of "right-hand rules": devices for remembering certain complex physical interactions by comparing the lines of movement or force to parts of the right hand. In the present context, a right-hand rule makes it easier to visualize the mutually perpendicular directions of electromagnetic waves, electric field, and magnetic field.

A field is a region of space in which it is possible to define the physical properties of each point in the region at any given moment in time. Thus, an electrical field and magnetic field are simply regions in which electrical and magnetic components, respectively, of electromagnetic force are exerted.

Hold out your right hand, palm perpendicular to the floor and thumb upright. Your fingers indicate the direction that an electromagnetic wave is moving. Your thumb points in the direction of the electrical field, as does the heel of your hand: the electrical field forms a plane perpendicular to the direction of wave propagation. Similarly, both your palm and the back of your hand indicate the direction of the magnetic field, which is perpendicular both to the electrical field and the direction of wave propagation.

The Electromagnetic Spectrum

As stated earlier, an electromagnetic wave is transverse, meaning that even as it moves forward, it oscillates in a direction perpendicular to the line of propagation. An electromagnetic wave can thus be defined as a transverse wave with mutually perpendicular electrical and magnetic fields that emanate from it.

The electromagnetic spectrum is the complete range of electromagnetic waves on a continuous distribution from a very low range of frequencies and energy levels, with a correspondingly long wavelength, to a very high range of frequencies and energy levels, with a correspondingly short wavelength. Included on the electromagnetic spectrum are radio waves and microwaves; infrared, visible, and ultraviolet light; x rays, and gamma rays. Though each occupies a definite place on the spectrum, the divisions between them are not firm: as befits the nature of a spectrum, one simply "blurs" into another.

FREQUENCY RANGE OF THE ELECTROMAGNETIC SPECTRUM. The range of frequencies for waves in the electromagnetic spectrum is from approximately 10^2 Hz to more than 10^{25} Hz. These numbers are an example of scientific notation, which makes it possible to write large numbers without having to include a string of zeroes. Without scientific notation, the large numbers used for discussing properties of the electromagnetic spectrum can become bewildering.

The first number given, for extremely low-frequency radio waves, is simple enough—100—but the second would be written as 1 followed by 25 zeroes. (A good rule of thumb for scientific notation is this: for any power n of 10, simply attach that number of zeroes to 1. Thus 10^6 is 1 followed by 6 zeroes, and so on.) In any case, 10^{25} is a much simpler figure than 10,000,000,000,000,000,000,000,000—or 10 trillion trillion. As noted earlier, gigahertz, or units of 1 billion Hertz, are often used in describing extremely high frequencies, in which case the number is written as 10^{16} GHz. For simplicity's sake, however, in the present context, the simple unit of Hertz (rather than kilo-, mega-, or gigahertz) is used wherever it is convenient to do so.

WAVELENGTHS ON THE ELECTROMAGNETIC SPECTRUM. The range of wavelengths found in the electromagnetic spectrum is from about 10^8 centimeters to less than 10^{-15} centimeters. The first number, equal to 1 million meters (about 621 mi), obviously expresses a great length. This figure is for radio waves of extremely low frequency; ordinary radio waves of the kind used for actual radio broadcasts are closer to 10^5 centimeters (about 328 ft).

For such large wavelengths, the use of centimeters might seem a bit cumbersome; but, as with the use of Hertz for frequencies, centimeters provide a simple unit that can be used to measure all wavelengths. Some charts of the electromagnetic spectrum nonetheless give figures in meters, but for parts of the spectrum beyond microwaves, this, too, can become challenging. The ultra-short wavelengths of gamma rays, after all, are equal to one-trillionth of a centimeter. By comparison, the angstrom—a unit so small it is used to measure the diameter of an atom—is 10 million times as large.

ENERGY LEVELS ON THE ELECTROMAGNETIC SPECTRUM. Finally, in terms of photon energy, the unit of measurement is the electron volt (eV), which is used for quantifying the energy in atomic particles. The range of photon energy in the electromagnetic spectrum is from about 10^{-13} to more than 10^{10} electron volts. Expressed in terms of joules, an electron volt is equal to $1.6 \cdot 10^{-19}$ J.

To equate these figures to ordinary language would require a lengthy digression; suffice it to say that even the highest ranges of the electromagnetic spectrum possess a small amount of energy in terms of joules. Remember, however, that the energy level identified is for a photon—a light particle. Again, without going into a great deal of detail, one can just imagine how many of these particles, which are much smaller than atoms, would fit into even the smallest of spaces. Given the fact that electromagnetic waves are traveling at a speed equal to that of light, the amount of photon energy transmitted in a single second is impressive, even for the lower ranges of the spectrum. Where gamma rays are concerned, the energy levels are positively staggering.

REAL-LIFE APPLICATIONS

The Radio Sub-Spectrum

Among the most familiar parts of the electromagnetic spectrum, in modern life at least, is radio. In most schematic representations of the spectrum, radio waves are shown either at the left end or the bottom, as an indication of the fact that these are the electromagnetic waves with the lowest frequencies, the longest wavelengths, and the smallest levels of photon energy. Included in this broad sub-spectrum, with frequencies up to about 10^7 Hertz, are long-wave radio, short-wave radio, and microwaves. The areas of communication affected are many: broadcast radio, television, mobile phones, radar—and even highly specific forms of technology such as baby monitors.

Though the work of Maxwell and Hertz was foundational to the harnessing of radio waves for human use, the practical use of radio had its beginnings with Marconi. During the 1890s, he made the first radio transmissions, and, by the end of the century, he had succeeded in transmitting telegraph messages across the Atlantic Ocean—a feat which earned him the Nobel Prize for physics in 1909.

Marconi's spark transmitters could send only coded messages, and due to the broad, long-wavelength signals used, only a few stations could broadcast at the same time. The development of the electron tube in the early years of the twentieth century, however, made it possible to transmit narrower signals on stable frequencies. This, in turn, enabled the development of technology for sending speech and music over the airwaves.

Broadcast Radio

THE DEVELOPMENT OF AM AND FM. A radio signal is simply a carrier: the process of adding information—that is, complex sounds such as those of speech or music—is called modulation. The first type of modulation developed was AM, or amplitude modulation, which Canadian-American physicist Reginald Aubrey Fessenden (1866-1932) demonstrated with the first United States radio broadcast in 1906. Amplitude modulation varies the instantaneous amplitude of the radio wave, a function of the radio station's power, as a means of transmitting information.

By the end of World War I, radio had emerged as a popular mode of communication: for the first time in history, entire nations could hear the same sounds at the same time. During the 1930s, radio became increasingly important, both for entertainment and information. Families in the era of the Great Depression would gather around large "cathedral radios"—so named for their size and shape—to hear comedy programs, soap operas, news programs, and speeches by important public figures such as President Franklin D. Roosevelt.

Throughout this era—indeed, for more than a half-century from the end of the first World War to the height of the Vietnam Conflict in the mid-1960s—AM held a dominant position in radio. This remained the case despite a number of limitations inherent in amplitude modulation: AM broadcasts flickered with popping noises from lightning, for instance, and cars with AM radios tended to lose their signal when going under a bridge. Yet, another mode of radio transmission was developed in the 1930s, thanks to American inventor and electrical engineer Edwin H. Armstrong (1890-1954). This was FM, or frequency modulation, which varied the radio signal's frequency rather than its amplitude.

Not only did FM offer a different type of modulation; it was on an entirely different frequency range. Whereas AM is an example of a long-wave radio transmission, FM is on the microwave sector of the electromagnetic spectrum, along with television and radar. Due to its high frequency and form of modulation, FM offered a "clean" sound as compared with AM. The addition of FM stereo broadcasts in the 1950s offered still further improvements; yet despite the advantages of FM, audiences were slow to change, and FM did not become popular until the mid- to late 1960s.

SIGNAL PROPAGATION. AM signals have much longer wavelengths, and smaller frequencies, than do FM signals, and this, in turn, affects the means by which AM signals are propagated. There are, of course, much longer radio wavelengths; hence, AM signals are described as intermediate in wavelength. These intermediate-wavelength signals reflect off highly charged layers in the ionosphere between 25 and 200 mi (40-332 km) above Earth's surface. Short-wavelength signals, such as those of FM, on the other hand, follow a straight-line path. As a result, AM broadcasts extend much farther than FM, particularly at night.

At a low level in the ionosphere is the D layer, created by the Sun when it is high in the sky. The D layer absorbs medium-wavelength signals during the day, and for this reason, AM signals do not travel far during daytime hours. After the Sun goes down, however, the D layer soon fades, and this makes it possible for AM signals to reflect off a much higher layer of the ionosphere known as the F layer. (This is also sometimes known as the Heaviside layer, or the Kennelly-Heaviside layer, after English physicist Oliver Heaviside and British-American electrical engineer Arthur Edwin Kennelly, who independently discovered the ionosphere in 1902.) AM signals "bounce" off the F layer as though it were a mirror, making it possible for a listener at night to pick up a signal from halfway across the country.

The Sun has other effects on long-wave and intermediate-wave radio transmissions. Sunspots, or dark areas that appear on the Sun in cycles of about 11 years, can result in a heavier buildup of the ionosphere than normal, thus impeding radio-signal propagation. In addition, occasional bombardment of Earth by charged particles from the Sun can also disrupt transmissions.

Due to the high frequencies of FM signals, these do not reflect off the ionosphere; instead, they are received as direct waves. For this reason, an FM station has a fairly short broadcast range, and this varies little with regard to day or night. The limited range of FM stations as compared to AM means that there is much less interference on the FM dial than for AM.

Distribution of Radio Frequencies

In the United States and most other countries, one cannot simply broadcast at will; the airwaves are regulated, and, in America, the governing authority is the Federal Communications Commission (FCC). The FCC, established in 1934, was an outgrowth of the Federal Radio Commission, founded by Congress seven years earlier. The FCC actually "sells air," charging companies a fee to gain rights to a certain frequency. Those companies may in turn sell that air to others for a profit.

At the time of the FCC's establishment, AM was widely used, and the federal government assigned AM stations the frequency range of 535 kHz to 1.7 MHz. Thus, if an AM station today is called, for instance, "AM 640," this means that it operates at 640 kHz on the dial. The FCC assigned the range of 5.9 to 26.1 MHz to shortwave radio, and later the area of 26.96 to 27.41 MHz to citizens' band (CB) radio. Above these are microwave regions assigned to television stations, as well as FM, which occupies the range from 88 to 108 MHz.

The organization of the electromagnetic spectrum's radio frequencies—which, of course, is an entirely arbitrary, humanmade process—is fascinating. It includes assigned frequencies for everything from garage-door openers to deep-space radio communications. The FCC recognizes seven divisions of radio carriers, using a system that is not so much based on rational rules as it is on the way that the communications industries happened to develop over time.

THE SEVEN FCC DIVISIONS. Most of what has so far been described falls under the heading of "Public Fixed Radio Services": AM and FM radio, other types of radio such as shortwave, television, various other forms of microwave broadcasting, satellite systems, and communication systems for federal departments and agencies. "Public Mobile Services" include pagers, air-to-ground service (for example, aircraft-to-tower communications), offshore service for sailing vessels, and rural radio-telephone service. "Commercial Mobile Radio Services" is the realm of cellular phones, and "Personal Communications Service" that of the newer wireless technology that began to challenge cellular for market dominance in the late 1990s.

"Private Land Mobile Radio Service" (PMR) and "Private Operational-Fixed Microwave Services" (OFS) are rather difficult to distinguish, the principal difference being that the former is used exclusively by profit-making businesses, and the latter mostly by nonprofit institutions. An example of PMR technology is the dispatching radios used by taxis, but this is only one of the more well-known forms of internal electronic communications for industry. For instance, when a film production company is shooting a picture and the director needs to speak to someone at the producer's trailer a mile away, she may use PMR radio technology. OFS was initially designated purely for nonprofit use, and is used often by schools; but banks and other profit-making institutions often use OFS because of its low cost.

Finally, there is the realm of "Personal Radio Services," created by the FCC in 1992. This branch, still in its infancy, will probably one day include video-on-demand, interactive polling, online shopping and banking, and other activities classified under the heading of Interactive Video and Data Services, or IVDS. Unlike other types of video technology, these will all be wireless, and, therefore, represent a telecommunications revolution all their own.

Microwaves

MICROWAVE COMMUNICATION. Though microwaves are treated separately from radio waves, in fact, they are just radio signals of a very short wavelength. As noted earlier, FM signals are actually carried on microwaves, and, as with FM in particular, microwave signals in general are very clear and very strong, but do not extend over a great geographical area. Nor does microwave include only high-frequency radio and television; in fact, any type of information that can be transmitted via telephone wires or coaxial cables can also be sent via a microwave circuit.

Microwaves have a very narrow, focused beam: thus, the signal is amplified considerably when an antenna receives it. This phenomenon, known as "high antenna gain," means that microwave transmitters need not be highly powerful to produce a strong signal. To further the reach of microwave broadcasts, transmitters are often placed atop mountain peaks, hilltops, or tall buildings. In the past, a microwave-transmitting network such as NBC (National Broadcasting Company) or CBS (Columbia Broadcasting System) required a network of ground-based relay stations to move its signal across the continent. The advent of satellite broadcasting in the 1960s, however, changed much about the way signals are beamed: today, networks typically replace, or at least augment, ground-based relays with satellite relays.

The first worldwide satellite TV broadcast, in the summer of 1967, featured the Beatles singing their latest song "All You Need Is Love." Due to the international character of the broadcast, with an estimated 200 million viewers, John Lennon and Paul McCartney wrote a song with simple, universal lyrics, and the result was just another example of electronic communication uniting large populations. Indeed, the phenomenon of rock music, and of superstardom as people know it today, would be impossible without many of the forms of technology discussed here. Long before the TV broadcast, the Beatles had come to fame through the playing of their music on the radio waves—and, thus, they owed much to Maxwell, Hertz, and Marconi.

MICROWAVE OVENS. The same microwaves that transmit FM and television signals—to name only the most obviously applications of microwave for communication—can also be harnessed to cook food. The microwave oven, introduced commercially in 1955, was an outgrowth of military technology developed a decade before.

During World War II, the Raytheon Manufacturing Company had experimented with a magnetron, a device for generating extremely short-wavelength radio signals as a means of improving the efficiency of military radar. While working with a magnetron, a technician named Percy Spencer was surprised to discover that a candy bar in his pocket had melted, even though he had not felt any heat. This led him to considering the possibilities of applying the magnetron to peacetime uses, and a decade later, Raytheon's "radar range" hit the market.

Those early microwave ovens had none of varied power settings to which modern users of the microwave—found today in two-thirds of all American homes—are accustomed. In the first microwaves, the only settings were "on" and "off," because there were only two possible adjustments: either the magnetron would produce, or not produce, microwaves. Today, it is possible to use a microwave for almost anything that involves the heating of food that contains water—from defrosting a steak to popping popcorn.

As noted much earlier, in the general discussion of electromagnetic radiation, there are three basic types of heat transfer: conduction, convection, and radiation. Without going into too much detail here, conduction generally involves heat transfer between molecules in a solid; convection takes place in a fluid (a gas such as air or a liquid such as water); and radiation, of course, requires no medium.

A conventional oven cooks through convection, though conduction also carries heat from the outer layers of a solid (for example, a turkey) to the interior. A microwave, on the other hand, uses radiation to heat the outer layers of the food; then conduction, as with a conventional oven, does the rest. The difference is that the microwave heats only the food—or, more specifically, the water, which then transfers heat throughout the item being heated—and not the dish or plate. Thus, many materials, as long as they do not contain water, can be placed in a microwave oven without being melted or burned. Metal, though it contains no water, is unsafe because the microwaves bounce off the metal surfaces, creating a microwave buildup that can produce sparks and damage the oven.

In a microwave oven, microwaves emitted by a small antenna are directed into the cooking compartment, and as they enter, they pass a set of turning metal fan blades. This is the stirrer, which disperses the microwaves uniformly over the surface of the food to be heated. As a microwave strikes a water molecule, resonance causes the molecule to align with the direction of the wave. An oscillating magnetron causes the microwaves to oscillate as well, and this, in turn, compels the water molecules to do the same. Thus, the water molecules are shifting in position several million times a second, and this vibration generates energy that heats the water.

Radio Waves for Measurement and Ranging

RADAR. Radio waves can be used to send communication signals, or even to cook food; they can also be used to find and measure things. One of the most obvious applications in this regard is radar, an acronym for *RA*dio *D*etection *A*nd *R*anging.

Radio makes it possible for pilots to "see" through clouds, rain, fog, and all manner of natural phenomena—not least of which is darkness. It can also identify objects, both natural and manmade, thus enabling a peacetime pilot to avoid hitting another craft or the side of a mountain. On the other hand, radar may help a pilot in wartime to detect the presence of an enemy. Nor is radar used only in the skies, or for military purposes, such as guiding missiles: on the ground, it is used to detect the speeds of objects such as automobiles on an interstate highway, as well as to track storms.

In the simplest model of radar operation, the unit sends out microwaves toward the target, and the waves bounce back off the target to the unit. Though the speed of light is reduced somewhat, due to the fact that waves are traveling through air rather than through a vacuum, it is, nonetheless, possible to account for this difference. Hence, the distance to the target can be calculated using the simple formula $d = vt$, where d is distance, v is velocity, and t is time.

Typically, a radar system includes the following: a frequency generator and a unit for controlling the timing of signals; a transmitter and, as with broadcast radio, a modulator; a duplexer, which switches back and forth between transmission and reception mode; an antenna; a receiver, which detects and amplifies the signals bounced back to the antenna; signal and data processing units; and data display units. In a monostatic unit—one in which the transmitter and receiver are in the same location—the unit has to be continually switched between sending and receiving modes. Clearly, a bistatic unit—one in which the transmitter and receiver antennas are at different locations—is generally preferable; but on an airplane, for instance, there is no choice but to use a monostatic unit.

In order to determine the range to a target—whether that target be a mountain, an enemy aircraft, or a storm—the target itself must first be detected. This can be challenging, because only a small portion of the transmitted pulse comes back to the receiving antenna. At the same time, the antenna receives reflections from a number of other objects, and it can be difficult to determine which signal comes from the target. For an aircraft in a wartime situation, these problems are compounded by the use of enemy countermeasures such as radar "jamming." Still another difficulty facing a military flyer is the fact that the use of radar itself—that is, the transmission of microwaves—makes the aircraft detectable to opposing forces.

TELEMETRY. Telemetry is the process of making measurements from a remote location and transmitting those measurements to receiving equipment. The earliest telemetry systems, developed in the United States during the 1880s, monitored the distribution and use of electricity in a given region, and relayed this information back to power companies using telephone lines. By the end of World War I, electric companies used the power lines themselves as information relays, and though such electrical telemetry systems remain in use in some sectors, most modern telemetry systems apply radio signals.

An example of a modern telemetry application is the use of an input device called a transducer to measure information concerning an astronaut's vital signs (heartbeat, blood pressure, body temperature, and so on) during a manned space flight. The transducer takes this information and converts it into an electrical impulse, which is then beamed to the space monitoring station on Earth. Because this signal carries information, it must be modulated, but there is little danger of interference with broadcast transmissions on Earth. Typically, signals from spacecraft are sent in a range above 10^{10} Hz, far above the frequencies of most microwave transmissions for commercial purposes.

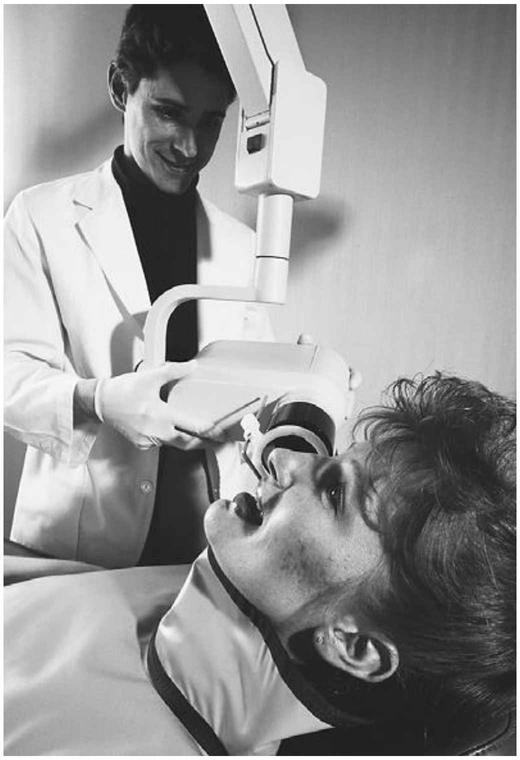
Light: Invisible, Visible, and Invisible Again

Between about 10^{13} and 10^{17} Hz on the electromagnetic spectrum is the range of light: infrared, visible, and ultraviolet. Light actually constitutes a small portion of the spectrum, and the area of visible light is very small indeed, extending from about $4.3 \cdot 10^{14}$ to $7.5 \cdot 10^{14}$ Hz. The latter, incidentally, is another example of scientific notation: not only is it easier not to use a string of zeroes, but where a coefficient or factor (for example, 4.3 or 7.5) is other than a multiple of 10, it is preferable to use what are called significant figures—usually a single digit followed by a decimal point and up to 3 decimal places.

Infrared light lies just below visible light in frequency, and this is easy to remember because of the name: red is the lowest in frequency of all the colors. Similarly, ultraviolet lies beyond the highest-frequency color, violet. Visible light itself, by far the most familiar part of the spectrum—especially prior to the age of radio communications—is discussed in detail elsewhere.

INFRARED LIGHT. Though we cannot see infrared light, we feel it as heat. German-English astronomer William Herschel (1738-1822), first scientist to detect infrared radiation from the Sun, demonstrated its existence in 1800 by using a thermometer. Holding a prism, a three-dimensional glass shape used for diffusing beams of light, he directed a beam of sunlight toward the thermometer, which registered the heat of the infrared rays.

Eighty years later, English scientist Sir William Abney (1843-1920) developed infrared photography, a method of capturing infrared radiation, rather than visible light, on film. By the mid-twentieth century, infrared photography had come into use for a variety of purposes. Military forces, for instance, may use infrared to



"Soft" x-ray machines, such as this one being used by a dentist to photograph the patient's teeth, operate at relatively low frequencies and thus don't harm the patient. *(Photograph by Richard T. Nowitz/Corbis. Reproduced by permission.)*

detect the presence of enemy troops. Medicine makes use of infrared photography for detecting tumors, and astronomers use infrared to detect stars too dim to be seen using ordinary visible light.

The uses of infrared imaging in astronomy, as a matter of fact, are many. The development in the 1980s of infrared arrays, two-dimensional grids which produce reliable images of infrared phenomena, revolutionized infrared astronomy. Because infrared penetrates dust much more easily than does visible light, infrared astronomy makes it easier to see regions of the universe where stars—formed from collapsing clouds of gas and dust—are in the process of developing. Because hydrogen molecules emit infrared radiation, infrared astronomy helps provide clues regarding the distribution of this highly significant chemical element throughout the universe.

ULTRAVIOLET LIGHT. Very little of the Sun's ultraviolet light penetrates Earth's atmosphere—a fortunate thing, since ultraviolet (UV) radiation can be very harmful to human skin. A suntan, as a matter of fact, is actually the skin's defense against these harmful UV rays. Due to the fact that Earth is largely opaque, or resistant, to ultraviolet light, the most significant technological applications of UV radiation are found in outer space.

In 1978 the United States, in cooperation with several European space agencies, launched the International Ultraviolet Explorer (IUE), which measured the UV radiation from tens of thousands of stars, nebulae, and galaxies. Despite the progress made with IUE, awareness of its limitations—including a mirror of only 17 in (45 cm) on the telescope itself—led to the development of a replacement in 1992.

This was the Extreme Ultraviolet Explorer (EUVE), which could observe UV phenomena over a much higher range of wavelengths than those observed by IUE. In addition, the Hubble Space Telescope, launched by the United States in 1990, includes a UV instrument called the Goddard High Resolution Spectrograph. With a mirror measuring 8.5 ft (2.6 m), it is capable of observing objects much more faint than those detected earlier by IUE.

Ultraviolet astronomy is used to study the winds created by hot stars, as well as stars still in the process of forming, and even stars that are dying. It is also useful for analyzing the densely packed, highly active sectors near the centers of galaxies, where both energy and temperatures are extremely high.

X Rays

Though they are much higher in frequency than visible light—with wavelengths about 1,000 times shorter than for ordinary light rays—x rays are a familiar part of modern life due to their uses in medicine. German scientist Wilhelm Röntgen (1845-1923) developed the first x-ray device in 1895, and, thus, the science of using x-ray machines is called roentgenology.

The new invention became a curiosity, with carnivals offering patrons an opportunity to look at the insides of their hands. And just as many people today fear the opportunities for invasion of privacy offered by computer technology, many at the time worried that x rays would allow robbers and peeping toms to look into people's houses. Soon, however, it became clear that the most important application of x rays lay in medicine.

KEY TERMS

AMPLITUDE: The maximum displacement of a vibrating material. In wave motion, amplitude is the "size" of a wave, an indicator of the energy and intensity of the wave.

CYCLE: One complete oscillation. In wave motion, this is equivalent to the movement of a wave from trough to crest and back to trough.

ELECTROMAGNETIC FORCE: The total force on an electrically charged particle, which is a combination of forces due to electrical and/or magnetic fields around the particle. Electromagnetic force reflects electromagnetic interaction, one of the four fundamental interactions in nature.

ELECTROMAGNETIC SPECTRUM: The complete range of electromagnetic waves on a continuous distribution from a very low range of frequencies and energy levels, with a correspondingly long wavelength, to a very high range of frequencies and energy levels, with a correspondingly short wavelength. Included on the electromagnetic spectrum are long-wave and short-wave radio; microwaves; infrared, visible, and ultraviolet light; x rays, and gamma rays.

ELECTROMAGNETIC WAVE: A transverse wave with electrical and magnetic fields that emanate from it. The directions of these fields are perpendicular to one another, and both are perpendicular to the line of propagation for the wave itself.

ELECTROMAGNETISM: The branch of physics devoted to the study of electrical and magnetic phenomena.

FIELD: A region of space in which it is possible to define the physical properties of each point in the region at any given moment in time.

FREQUENCY: In wave motion, frequency is the number of waves passing through a given point during the interval of one second. The higher the frequency, the shorter the wavelength. Measured in Hertz, frequency is mathematically related to wave speed, wavelength, and period.

FUNDAMENTAL INTERACTION: The basic mode by which particles interact. There are four known fundamental interactions in nature: gravitational, electromagnetic, strong nuclear, and weak nuclear.

HARMONIC MOTION: The repeated movement of a particle about a position of equilibrium, or balance.

HERTZ: A unit for measuring frequency, named after nineteenth-century German physicist Heinrich Rudolf Hertz (1857-1894). High frequencies are expressed in terms of kilohertz (kHz; 10^3 or 1,000 cycles per second); megahertz (MHz; 10^6 or 1 million cycles per second); and gigahertz (GHz; 10^9 or 1 billion cycles per second.)

INTENSITY: Intensity is the rate at which a wave moves energy per unit of cross-sectional area.

OSCILLATION: A type of harmonic motion, typically periodic, in one or more dimensions.

PERIOD: For wave motion, a period is the amount of time required to complete one full cycle. Period is mathematically related to frequency, wavelength, and wave speed.

KEY TERMS CONTINUED

PERIODIC MOTION: Motion that is repeated at regular intervals. These intervals are known as periods.

PERIODIC WAVE: A wave in which a uniform series of crests and troughs follow one after the other in regular succession.

PHOTON: A particle of electromagnetic radiation carrying a specific amount of energy, measured in electron volts (eV). For parts of the electromagnetic spectrum with a low frequency and long wavelength, photon energy is relatively low; but for parts with a high frequency and short wavelength, the value of photon energy is very high.

PROPAGATION: The act or state of traveling from one place to another.

RADIATION: The transfer of energy by means of electromagnetic waves, which require no physical medium (for example, water or air) for the transfer. Earth receives the Sun's energy, via the electromagnetic spectrum, by means of radiation.

SCIENTIFIC NOTATION: A method used by scientists for writing extremely large numbers. This usually involves a coefficient, or factor, of a single digit followed by a decimal point and up to three decimal places, multiplied by 10 to a given exponent. Thus, instead of writing 75,120,000, the preferred scientific notation is $7.512 \cdot 10^7$. To visualize the value of very large multiples of 10, it is helpful to remember that the value of 10 raised to any power n is the same as 1 followed by that number of zeroes. Hence 10^{25}, for instance, is simply 1 followed by 25 zeroes.

TRANSVERSE WAVE: A wave in which the vibration or motion is perpendicular to the direction in which the wave is moving.

WAVELENGTH: The distance between a crest and the adjacent crest, or the trough and an adjacent trough, of a wave. Wavelength, symbolized λ (the Greek letter lambda) is mathematically related to wave speed, period, and frequency.

WAVE MOTION: A type of harmonic motion that carries energy from one place to another without actually moving any matter.

HOW A MEDICAL X-RAY MACHINE WORKS. Due to their very short wavelengths, x rays can pass through substances of low density—for example, fat and other forms of soft tissue—without their movement being interrupted. But in materials of higher density, such as bone, atoms are packed closely together, and this provides x rays with less space through which to travel. As a result, x-ray images show dark areas where the rays traveled completely through the target, and light images of dense materials that blocked the movement of the rays.

Medical x-ray machines are typically referred to either as "hard" or "soft." Soft x rays are the ones with which most people are more familiar. Operating at a relatively low frequency, these are used to photograph bones and internal organs, and provided the patient does not receive prolonged exposure to the rays, they cause little damage. Hard x rays, on the other hand, are designed precisely to cause damage—not to the patient, but to cancer cells. Because they use high voltage and high-frequency rays, hard x rays can be quite dangerous to the patient as well.

OTHER APPLICATIONS. X-ray crystallography, developed in the early twentieth century, is devoted to the study of the interference patterns produced by x rays passing through materials that are crystalline in Structure. Each of these discoveries, in turn, transformed daily life: insulin, by offering hope to diabetics, penicillin, by providing a treatment for a number of previously fatal illnesses, and DNA, by enabling scientists to make complex assessments of genetic information.

In addition to the medical applications, the scanning capabilities of x-ray machines make them useful for security. A healthy person receives an x ray at a doctor's office only once in a while; but everyone who carries items past a certain point in a major airport must submit to x-ray security scanning. If one is carrying a purse or briefcase, for instance, this is placed on a moving belt and subjected to scanning by a low-power device that can reveal the contents.

Gamma Rays

At the furthest known reaches of the electromagnetic spectrum are gamma rays, ultra high-frequency, high-energy, and short- wavelength forms of radiation. Human understanding of gamma rays, including the awesome powers they contain, is still in its infancy.

In 1979, a wave of enormous energy passed over the Solar System. Though its effects on Earth were negligible, instruments aboard several satellites provided data concerning an enormous quantity of radiation caused by gamma rays. As to the source of the rays themselves, believed to be a product of nuclear fusion on some other body in the universe, scientists knew nothing.

The Compton Gamma Ray Observatory Satellite, launched by NASA (National Aeronautics and Space Administration) in 1991, detected a number of gamma-ray bursts over the next two years. The energy in these bursts was staggering: just one of these, scientists calculated, contained more than a thousand times as much energy as the Sun will generate in its entire lifetime of 10 billion years.

Some astronomers speculate that the source of these gamma-ray bursts may ultimately be a distant supernova, or exploding star. If this is the case, scientists may have found the supernova; but do not expect to see it in the night sky. It is not known just how long ago it exploded, but its light appeared on Earth some 340,000 years ago, and during that time it was visible in daylight for more than two years. So great was its power that the effects of this stellar phenomenon are still being experienced.

WHERE TO LEARN MORE

Branley, Franklyn Mansfield. *The Electromagnetic Spectrum: Key to the Universe.* Illustrated by Leonard D. Dank. New York: Crowell, 1979.

Electromagnetic Spectrum (Web site). <http://www.jsc.mil/images/speccht.jpg> (April 25, 2001).

The Electromagnetic Spectrum (Web site). <http://library.thinkquest.org/11119/index.htm> (April 30, 2001).

The Electromagnetic Spectrum (Web site). <http://www.ospi.wednet.edu:8001/curric/space/lfs/emspectr.html> (April 30, 2001).

Electromagnetic Spectrum/NASA: National Aeronautics and Space Administration (Web site). <http://imagine.gsfc.nasa.gov/docs/science/know_11/emspectrum.html> (April 30, 2001).

"How the Radio Spectrum Works." How Stuff Works (Web site). <http://www.howstuffworks.com/radio~spectrum.htm> (April 25, 2001).

Internet Resources for Sound and Light (Web site). <http://electro.sau.edu/SLResources.html> (April 25, 2001).

Nassau, Kurt. *Experimenting with Color.* New York: F. Watts, 1997.

"Radio Electronics Pages" ePanorama.net (Web site). <http://www.epanorama.net/radio.html> (April 25, 2001).

Skurzynski, Gloria. *Waves: The Electromagnetic Universe.* Washington, D.C.: National Geographic Society, 1996.

LIGHT

CONCEPT

Light exists along a relatively narrow bandwidth of the electromagnetic spectrum, and the region of visible light is more narrow still. Yet, within that realm are an almost infinite array of hues that quite literally give color to the entire world of human experience. Light, of course, is more than color: it is energy, which travels at incredible speeds throughout the universe. From prehistoric times, humans harnessed light's power through fire, and later, through the invention of illumination devices such as candles and gas lamps. In the late nineteenth century, the first electric-powered forms of light were invented, which created a revolution in human existence. Today, the power of lasers, highly focused beams of high-intensity light, make possible a number of technologies used in everything from surgery to entertainment.

HOW IT WORKS

Early Progress in Understanding of Light

The first useful observations concerning light came from ancient Greece. The Greeks recognized that light travels through air in rays, a term from geometry describing that part of a straight line that extends in one direction only. Upon entering some denser medium, such as glass or water, as Greek scientists noticed, the ray experiences refraction, or bending. Another type of incidence, or contact, between a light ray and any surface, is reflection, whereby a light ray returns, rather than being absorbed at the interface.

The Greeks worked out the basic laws governing reflection and refraction, observing, for instance, that in reflection, the angle of incidence is approximately equal to the angle of reflection. Unfortunately, they also subscribed to the erroneous concept of intromission—the belief that light rays originate in the eye and travel toward objects, making them visible. Some 1,500 years after the high point of Greek civilization, Arab physicist Alhasen (Ibn al-Haytham; c. 965-1039), sometimes called the greatest scientist of the Middle Ages, showed that light comes from a source such as the Sun, and reflects from an object to the eyes.

The next great era of progress in studies of light began with the Renaissance (c. 1300-c. 1600.) However, the most profound scientific achievements in this area belonged not to scientists, but to painters, who were fascinated by color, shading, shadows, and other properties of light. During the early seventeenth century, Galileo Galilei (1564-1642) and German astronomer Johannes Kepler (1571-1630) built the first refracting telescopes, while Dutch physicist and mathematician Willebrord Snell (1580-1626) further refined the laws of refraction.

The Spectrum

Sir Isaac Newton (1642-1727) was as intrigued with light as he was with gravity and the other concepts associated with his work. Though it was not as epochal as his contributions to mechanics, Newton's work in optics, an area of physics that studies the production and propagation of light, was certainly significant.

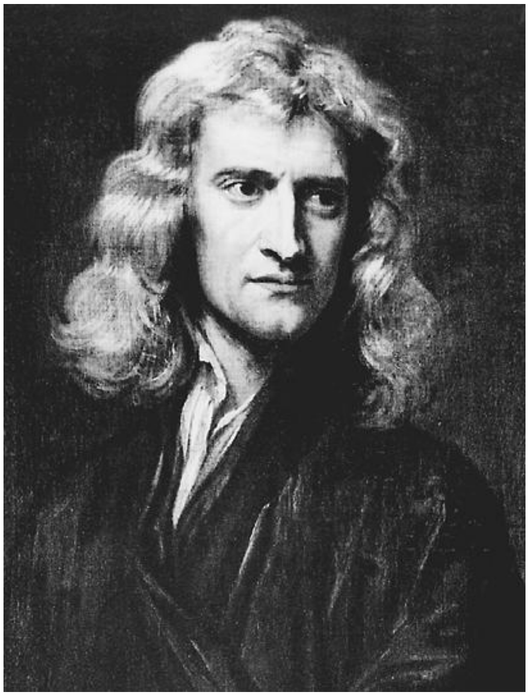
In Newton's time, physicists understood that a prism could be used for the diffusion of light rays—in particular, to produce an array of colors from a beam of white light. The prevailing belief was that white was a single color like the others, but Newton maintained that it was a combination of all other colors. To prove this, he directed a beam of white light through a prism, then allowed the diffused colors to enter another prism, at which point they recombined as white light.

Newton gave to the array of colors in visible light the term spectrum, (plural, "spectra") meaning the continuous distribution of properties in an ordered arrangement across an unbroken range. The term can be used for any set of characteristics for which there is a gradation, as opposed to an excluded middle. An ordinary light switch provides an example of a situation in which there is an excluded middle: there is nothing between "on" and "off." A dimmer switch, on the other hand, is a spectrum, because a very large number of gradations exist between the two extremes represented by a light switch.

SEVEN COLORS...OR SIX? The distribution of colors across the spectrum is as follows: red-orange-yellow-green-blue-violet. The reasons for this arrangement, explained below in the context of the electromagnetic spectrum, were unknown to Newton. Not only did he live in an age that had almost no understanding of electromagnetism, but he was also a product of the era called the Enlightenment, when intellectuals (scientists included) viewed the world as a highly rational, ordered mechanism. His Enlightenment viewpoint undoubtedly influenced his interpretation of the spectrum as a set of seven colors, just as there are seven notes on the musical scale.

In addition to the six basic colors listed above, Newton identified a seventh, indigo, between blue and violet. In fact, there is a noticeable band of color between blue and violet, but this is because one color fades into another. With a spectrum, there is a blurring of lines between one color and the next: for instance, orange exists at a certain point along the spectrum, as does yellow, but between them is a nearly unlimited number of orange-yellow and yellow-orange gradations.

Indigo itself is not really a distinct color—just a deep, purplish blue. But its inclusion in the listing of colors on the spectrum has given generations of students a handy mnemonic (memorization) device: the name "ROY G. BIV." These letters form an acrostic (a word constructed from the first letters of other words) for the colors of the spectrum. Incidentally, there is something arbitrary even in the idea of six colors, or for that matter seven musical notes: in both cases, there exists a very large gradation of shades, yet also in both cases, the divisions used were chosen for practical purposes.



Isaac Newton. *(The Bettmann Archive. Reproduced by permission.)*

Waves, Particles, and Other Questions Concerning Light

THE WAVE-PARTICLE CONTROVERSY BEGINS. Newton subscribed to the corpuscular theory of light: the idea that light travels as a stream of particles. On the other hand, Dutch physicist and astronomer Christiaan Huygens (1629-1695) maintained that light travels in waves. During the century that followed, adherents of particle theory did intellectual battle with proponents of wave theory. "Battle" is not too strong a word, because the conflict was heated, and had a nationalistic element. Reflecting both the burgeoning awareness of the nation-state among Europeans, as well as Britons' sense of their own island as an entity separate from the European continent, particle theory had its strongest defenders in Newton's homeland, while continental scientists generally accepted wave theory.

According to Huygens, the appearance of the spectrum, as well as the phenomena of reflection and refraction, indicated that light was a wave. Newton responded by furnishing complex mathematical calculations which showed that particles could exhibit the behaviors of reflection and refraction as well. Furthermore, Newton challenged, if light were really a wave, it should be able to bend around corners. Yet, in 1660, an experiment by Italian physicist Francesco Grimaldi (1618-1663) proved that light could do just that. Passing a beam of light through a narrow aperture, or opening, Grimaldi observed a phenomenon called diffraction, or the bending of light.

In view of the nationalistic character that the wave-particle debate assumed, it was ironic that the physicist whose work struck a particularly forceful blow against corpuscular theory was himself an Englishman: Thomas Young (1773-1829), who in 1801 demonstrated interference in light. Directing a light beam through two closely spaced pinholes onto a screen, Young reasoned that if light truly were made of particles, the beams would project two distinct points onto the screen. Instead, what he saw was a pattern of interference—a wave phenomenon.

THE QUESTION OF A MEDIUM. As the nineteenth century progressed, evidence in favor of wave theory grew. Experiments in 1850 by Jean Bernard Leon Foucault (1819-1868)—famous for his pendulum —showed that light traveled faster in air than through water. Based on studies of wave motion up to that time, Foucault's work added substance to the view of light as a wave.

Foucault also measured the speed of light in a vacuum, a speed which he calculated to within 1% of its value as it is known today: 186,000 mi (299,339 km) per second. An understanding of just how fast light traveled, however, caused a nagging question dating back to the days of Newton and Huygens to resurface: how did light travel?

All types of waves known to that time traveled through some sort of medium: for instance, sound waves were propagated through air, water, or some other type of matter. If light was a wave, as Huygens said, then it, too, must have some medium. Huygens and his followers proposed a weak theory by suggesting the existence of an invisible substance called ether, which existed throughout the universe and which carried light.

Ether, of course, was really no answer at all. There was no evidence that it existed, and to many scientists, it was merely a concept invented to shore up an otherwise convincing argument. Then, in 1872, Scottish physicist James Clerk Maxwell (1831-1879) proposed a solution that must have surprised many scientists. The "medium" through which light travels, Maxwell proposed, was no medium at all; rather, the energy in light is transferred by means of radiation, which requires no medium.

Electromagnetism

Maxwell brought together a number of concepts developed by his predecessors, sorting these out and adding to them. His work led to the identification of a "new" fundamental interaction, in addition to that associated with gravity. This was the mode of particle interaction associated with electromagnetic force.

The particulars of electromagnetic force, waves, and radiation are a subject unto themselves—really, many subjects. As for the electromagnetic spectrum, it is treated at some length in an essay elsewhere in this volume, and the reader is encouraged to review that essay to gain a greater understanding of light and its place in the spectrum.

In addition, some awareness of wave motion and related phenomena would also be of great value, and, for this purpose, other essays are recommended. In the present context, a number of topics relating to these larger subjects will be handled in short order, with a minimum of explanation, to enable a more speedy transition to the subject of principal importance here: light.

ELECTROMAGNETIC WAVES. There is, of course, no obvious connection between light and the electromagnetic force observed in electrical and magnetic interactions. Yet, light is an example of an electromagnetic wave, and is part of the electromagnetic spectrum. The breakthrough in establishing the electromagnetic quality of light can be attributed both to Maxwell and German physicist Heinrich Rudolf Hertz (1857-1894).

In his *Electricity and Magnetism* (1873), Maxwell suggested that electromagnetic force might have aspects of a wave phenomenon, and his experiments indicated that electromagnetic waves should travel at exactly the same speed as light. This appeared to be more than just a coincidence, and his findings led him to theorize that the electromagnetic interaction included not only electricity and magnetism, but light as well. Some time later, Hertz proved Maxwell's hypothesis by showing that electromagnetic waves obeyed the same laws of reflection, refraction, and diffraction as light.

Hertz also discovered the photoelectric effect, the process by which certain metals acquire an electrical potential when exposed to light. He could not explain this behavior, and, indeed, there was nothing in wave theory that could account for it. Strangely, after more than a century in which acceptance of wave theory had grown, he had encountered something that apparently supported what Newton had said long before: that light traveled in particles rather than waves.

THE WAVE-PARTICLE DEBATE REVISITED

One of the modern physicists whose name is most closely associated with the subject of light is Albert Einstein (1879-1955). In the course of proving that matter is convertible to energy, as he did with the theory of relativity, Einstein predicted that this could be illustrated by accelerating to speeds close to that of light. (Conversely, he also showed that it is impossible for matter to reach the speed of light, because to do so would—as he proved mathematically—result in the matter acquiring an infinite amount of mass, which, of course, is impossible.)

Much of Einstein's work was influenced by that of German physicist Max Planck (1858-1947), father of quantum theory. Quantum theory and quantum mechanics are, of course, far too complicated to explain in any depth here. It is enough to say that they called into question everything physicists thought they knew, based on Newton's theories of classical mechanics. In particular, quantum mechanics showed that, at the subatomic level, particles behave in ways not just different from, but opposite to, the behavior of larger physical objects in the observable world. When a quantity is "quantized," its values or properties at the atomic or subatomic level are separate from one another—meaning that something can both be one thing and its opposite, depending on how it is viewed.

Interpreting Planck's observations, Einstein in a 1905 paper on the photoelectric effect maintained that light is quantized—that it appears in "bundles" of energy that have characteristics both of waves and of particles. Though light travels in waves, as Einstein showed, these waves sometimes behave as particles, which is the case with the photoelectric effect. Nearly two decades later, American physicist Arthur Holly Compton (1892-1962) confirmed Einstein's findings and gave a name to the "particles" of light: photons.

LIGHT'S PLACE IN THE ELECTROMAGNETIC SPECTRUM

The electromagnetic spectrum is the complete range of electromagnetic waves on a continuous distribution from a very low range of frequencies and energy levels, with a correspondingly long wavelength, to a very high range of frequencies and energy levels, with a correspondingly short wavelength. Included on the electromagnetic spectrum are radio waves and microwaves; infrared, visible, and ultraviolet light; x rays, and gamma rays. As discussed earlier, concerning the visible color spectrum, each of these occupies a definite place on the spectrum, but the divisions between them are not firm: in keeping with the nature of a spectrum, one band simply "blurs" into another.

Of principal concern here is an area near the middle of the electromagnetic spectrum. Actually, the very middle of the spectrum lies within the broad area of infrared light, which has frequencies ranging from 10^{12} to just over 10^{14} Hz, with wavelengths of approximately 10^{-1} to 10^{-3} centimeters. Even at this point, the light waves are oscillating at a rate between 1 and 100 trillion times a second, and the wavelengths are from 1 millimeter to 0.01 millimeters. Yet, over the breadth of the electromagnetic spectrum, wavelengths get much shorter, and frequencies much greater.

Infrared lies just below visible light in frequency, which is easy to remember because of the name: red is the lowest in frequency of all the colors, as discussed below. Similarly, ultraviolet lies beyond the highest-frequency color, violet. Neither infrared nor ultraviolet can be seen, yet we experience them as heat. In the case of ultraviolet (UV) light, the rays are so powerful that exposure to even the minuscule levels of UV radiation that enter Earth's atmosphere can cause skin cancer.

Ultraviolet light occupies a much narrower band than infrared, in the area of about 10^{15} to 10^{16} Hz—in other words, oscillations between 1 and 10 quadrillion times a second. Wavelengths in this region are from just above 10^{-6} to about 10^{-7} centimeters. These are often measured in terms of a nanometer (nm)—equal to one-millionth of a millimeter—meaning that the wavelength range is from above 100 down to about 10 nm.

Between infrared and ultraviolet light is the region of visible light: the six colors that make up much of the world we know. Each has a specific range and frequency, and together they occupy an extremely narrow band of the electromagnetic spectrum: from $4.3 \cdot 10^{14}$ to $7.5 \cdot 10^{14}$ Hz in frequency, and from 700 down to 400 nm in wavelength. To compare its frequency range to that of the entire spectrum, for instance, is the same as comparing 3.2 to 100 billion.

REAL-LIFE APPLICATIONS

Colors

Unlike many of the topics addressed by physics, color is far from abstract. Numerous expressions in daily life describe the relationship between energy and color: "red hot," for instance, or "blue with cold." In fact, however, red—with a smaller frequency and a longer wavelength than blue—actually has less energy; therefore, blue objects should be hotter.

The phenomenon of the red shift, discovered in 1923 by American astronomer Edwin Hubble (1889-1953), provides a clue to this apparent contradiction. As Hubble observed, the light waves from distant galaxies are shifted to the red end, and he reasoned that this must mean those galaxies are moving away from the Milky Way, the galaxy in which Earth is located.

To generalize from what Hubble observed, when something shows red, it is moving away from the observer. The laws of thermodynamics state that where heat is involved, the movement is always away from an area of high temperature and toward an area of low temperature. Heated molecules that reflect red light are, thus, to use a colloquialism, "showing their tail end" as they move toward an area of low temperature. By contrast, molecules of low temperature reflect bluish or purple light because the tendency of heat is to move toward them.

There are other reasons, aside from heat, that some objects tend to be red and others blue—or another color. Chemical factors may be involved: atoms of neon, for example, can be made to vibrate at a particular wavelength, producing a specific color. In any case, the color that an object reflects is precisely the color that it does not absorb: thus, if something is red, that means it has absorbed every color of the spectrum but red.

WHY IS THE SKY BLUE? The placement of colors on the electromagnetic spectrum provides an answer to that age-old question posed by generations of children to their parents: "Why is the sky blue?" Electromagnetic radiation is scattered as it enters the atmosphere, but all forms of radiation are not scattered equally. Those having shorter wavelengths—that is, toward the blue end of the spectrum—tend to scatter more than those with longer wavelengths, on the red and orange end.

Yet the longer-wavelength light becomes visible at sunset, when the Sun's light enters the atmosphere at an angle. In addition, the dim quality of evening light means that it is easiest to see light of longer wavelengths. This effect is known as Rayleigh scattering, after English physicist John William Strutt, Lord Rayleigh (1842-1919), who discovered it in 1871. Thanks to Rayleigh's discovery, there is an explanation not only for the question of why the sky is blue, but why sunsets are red, orange, and gold.

RAINBOWS. On the subject of color as children perceive it, many a child has been fascinated by a rainbow, seeing in them something magical. It is easy to understand why children perceive these beautiful phenomena this way, and why people have invented stories such as that of the pot of gold at the end of a rainbow. In fact, a rainbow, like many other "magical" aspects of daily life, can be explained in terms of physics.

A rainbow, in fact, is simply an illustration of the visible light spectrum. Rain drops perform the role of tiny prisms, dispersing white sunlight, much as scientists before Newton had learned to do. But if there is a pot of gold at the end of the rainbow, it would be impossible to find. In order for a rainbow to be seen, it must be viewed from a specific perspective: the observer must be in a position between the sunlight and the raindrops.

Sunlight strikes raindrops in such a way that they are refracted, then reflected back at an angle so that they represent the entire visible light spectrum. Though they are beautiful to see, rainbows are neither magical nor impossible to reproduce artificially. Such rainbows can be produced, for instance, in the spiral of small water droplets emerging from a water hose, viewed when one's back is to the Sun.

PERCEPTION OF LIGHT AND COLOR

People literally live and die for colors: the colors of a flag, for instance, present a rallying point for soldiers, and different colors are assigned specific political meanings. Blue, both in the American and French flags, typically stands for liberty. Red can symbolize the blood shed by patriots, or it can mean some version of fraternity or brotherhood. Such is the case with the red of the French tricolor (red-white-blue); likewise, the red in the flag of the former Soviet Union and other Communist countries stood for the alleged international brotherhood of all working peoples. In Islamic countries, by contrast, green stands for the unity of all Muslims.

These are just a few examples, drawn from a specific realm—politics—illustrating the meanings that people ascribe to colors. Similarly, people find meanings in images presented to them by light itself. In his *Republic,* the ancient Greek philosopher Plato (c. 427-347 B.C.) offered a complex parable, intended to illustrate the difference between reality and illusion, concerning a group of slaves who do not recognize the difference between sunlight and the light of a torch in a cave. Modern writers have noted the similarities between Plato's cave and a phenomenon which the ancient philosopher could hardly have imagined: a movie theatre, in which an artificial light projects images—images that people sometimes perceive as being all too real—onto a screen.

People refer to "tricks of the light," as, for instance, when one seems to see an image in a fire. One particularly well-known "trick of the light," a mirage, is discussed below, but there are also manmade illusions created by light, shapes, and images. An optical illusion is something that produces a false impression in the brain, causing one to believe that something is as it appears, when, in fact, it is not. When two lines of equal length are placed side by side, but one has arrows pointed outward at either end while the other line has arrows pointing inward, it appears that the line with the inward-pointing arrows are shorter.

This is an example of the ways in which human perception plays a role in what people see. That topic, of course, goes far beyond physics and into the realms of psychology and the social sciences. Nonetheless, it is worthwhile to consider, from a physical standpoint, how humans see what they see—and sometimes see things that are not there.

A MIRAGE. Because they can be demonstrated in light waves as well as in sound waves, diffraction and interference are discussed in separate essays. As for refraction, or the bending of light waves, this phenomenon can be seen in the familiar example of a mirage. While driving down a road on a hot day, one may observe that there are pools of water up ahead, but by the time one approaches them, they disappear.

Of course, the pools were never there; light itself has created an optical illusion of sorts. As light moves from one material to another, it bends with a different angle of refraction, and, though, in this instance, it is traveling entirely through air, it is moving through regions of differing temperature. Light waves travel faster through warm air than through cool air, and, thus, when the light enters the area over the heated surface of the asphalt, it experiences refraction. The waves are thus bent, creating the impression of a reflection, which suggests to the observer that there is water up ahead.

HOW THE EYE SEES COLOR. White, as noted earlier, is the combination of all colors; black is the absence of color. Where ink, dye, or other forms of artificial pigmentation are concerned, of course, black is a "real" color, but in terms of light, it is not. In the same way, the experience of coldness is real, yet "cold" does not exist as a physical phenomenon: it is simply the absence of heat.

The mixture of pigmentation is an entirely different matter from the mixture of light. In artificial pigmentation, the primary colors—the three colors which, when mixed, yield the remainder of the shades on the rainbow—are red, blue, and yellow. Red mixed with blue creates purple, blue mixed with yellow makes green, and red mixed with yellow yields orange. Black and white are usually created by using natural substances of that color—chalk for white, for instance, or various oxides for black. For light, on the other hand, blue and red are primary colors, but the third primary color is green, not yellow. From these three primary colors, all other shades of the visible spectrum can be made.

The mechanism of the human eye responds to the three primary colors of the visible light spectrum: thus, the eye's retina is equipped with tiny cones that respond to red, blue, and green light. The cones respond to bright light; other structures called rods respond to dim light, and the pupil regulates the amount of light that enters the eye.

The eye responds with maximum sensitivity to light at the middle of the visible color spectrum—specifically, green light with a wavelength of about 555 nm. The optimal wavelength for maximum sensitivity in dim light is around 510 nm, on the blue end. It is difficult for the eye to recognize red light, at the far end of the spectrum, against a dark background. However, this can be an advantage in situations of relative darkness, which is why red light is often used to maintain vision for sailors, amateur astronomers, and the military on night maneuvers. Because there is not much difference between the darkness and the red light, the eye adjusts and is able to see beyond the red light into the darkness. A bright yellow or white light in such situations, on the other hand, would minimize visibility in areas beyond the light.

Artificial Light

PREHISTORIC LIGHTING TECHNOLOGY. Prehistoric humans did not know it, but they were making use of electromagnetic radiation when they lit and warmed their caves with light from a fire. Though it would seem that warmth was more essential to human survival than artificial light, in fact, it is likely that both functions emerged at about the same time: once humans began using fire for warmth, it would have been a relatively short time before they comprehended the power of fire to drive out both darkness and the fierce creatures (for instance, bears) that came with it.

These distant forebears advanced to the fashioning of portable lighting technology in the form of torches or rudimentary lamps. Torches were probably made by binding together resinous material from trees, while lamps were made either from stones with natural depressions, or from soft rocks—for example, soapstone or steatite—into which depressions were carved by using harder material. Most of the many hundreds of lamps found by archaeologists at sites in southwestern France are made of either limestone or sandstone. Limestone was a particularly good choice, since it conducts heat poorly; lamps made of sandstone, a good conductor of heat, usually had carved handles to protect the hands of the user.

ARTIFICIAL LIGHT IN PREMODERN TIMES. The history of lighting is generally divided into four periods, each of which overlap, and which together illustrate the slow pace of change in illumination technology. First was the primitive, a period encompassing the torches and lamps of prehistoric human beings—though, in fact, French peasants used the same lighting methods depicted in nearby cave paintings until World War I.

Next came the classical stage, the world of Greece and Rome. Earlier civilizations, such as that of Egypt, belong to the primitive era in lighting—before the relatively widespread adoption of the candle and of vegetable oil as fuel. Third was the medieval stage, which saw the development of metal lamps. Last came the modern or invention stage, which began with the creation of the glass lantern chimney by Leonardo da Vinci (1452-1519) in 1490, culminated with Thomas Edison's (1847-1931) first practical incandescent bulb in 1879, and continues today.

At various times, ancient peoples used the fat of seals, horses, cattle, and fish as fuel for lamps. (Whale oil, by contrast, entered widespread use only during the nineteenth century.) Primitive humans sometimes used entire animals—for example, the storm petrel, a bird heavy in fat—to provide light. Even without such cruel excesses, however, animal fat made for a smoky, dangerous, foul-smelling fire.

The use of vegetable oils, a much more efficient medium for lighting, did not take hold until Greek, and especially, Roman times. Animal oils remained in use, however, among the poor, whose homes often reeked with the odor of castor oil or fish oil. Because virtually all fuels came from edible sources, times of famine usually meant times of darkness as well.

The candle, as well as the use of vegetable oils, dates back to earliest antiquity, but the use of candles only became common among the richest citizens of Rome. Because it used animal fat, the candle was apparently a return to an earlier stage, but its hardened tallow actually represented a much safer, more stable fuel than lamp oil.

INCANDESCENT LIGHT. Lighting technology in the period from about 1500 to the late nineteenth century involved a number of improvements, but in one respect, little had progressed since prehistoric times: people were still burning fuel to provide illumination. This all changed with the invention of the incandescent bulb, which, though it is credited to Edison, was the product of experimentation that took place throughout the nineteenth century. As early as 1802, British scientist Sir Humphry Davy (1778-1829) showed that electricity running through thin strips of metal could heat them enough to cause them to give off light—that is, electromagnetic radiation.

Edison, in fact, was just one of several inventors in the 1870s attempting to develop a practical incandescent lamp. His innovation lay in his understanding of the parameters necessary for developing such a lamp—in particular, decreasing the electrical resistance in the lamp filament (the part that is heated) so that less energy would be required to light it. On October 19, 1879, using low-resistance filaments of carbon or platinum, combined with a high-resistance carbon filament in a vacuum-sealed glass container, Edison produced the first practical lightbulb.

Much has changed in the design of lightbulbs during the decades following Edison's ingenious invention, of course, but his design provided the foundation. There is just one problem with incandescent light, however—a problem inherent in the definition and derivation of the word incandescent, which comes from a Latin root meaning "to become hot." The efficiency of a light is determined by the ratio of light, or usable energy, to heat—which, except in the case of a campfire, is typically not a desirable form of energy where lighting is concerned.

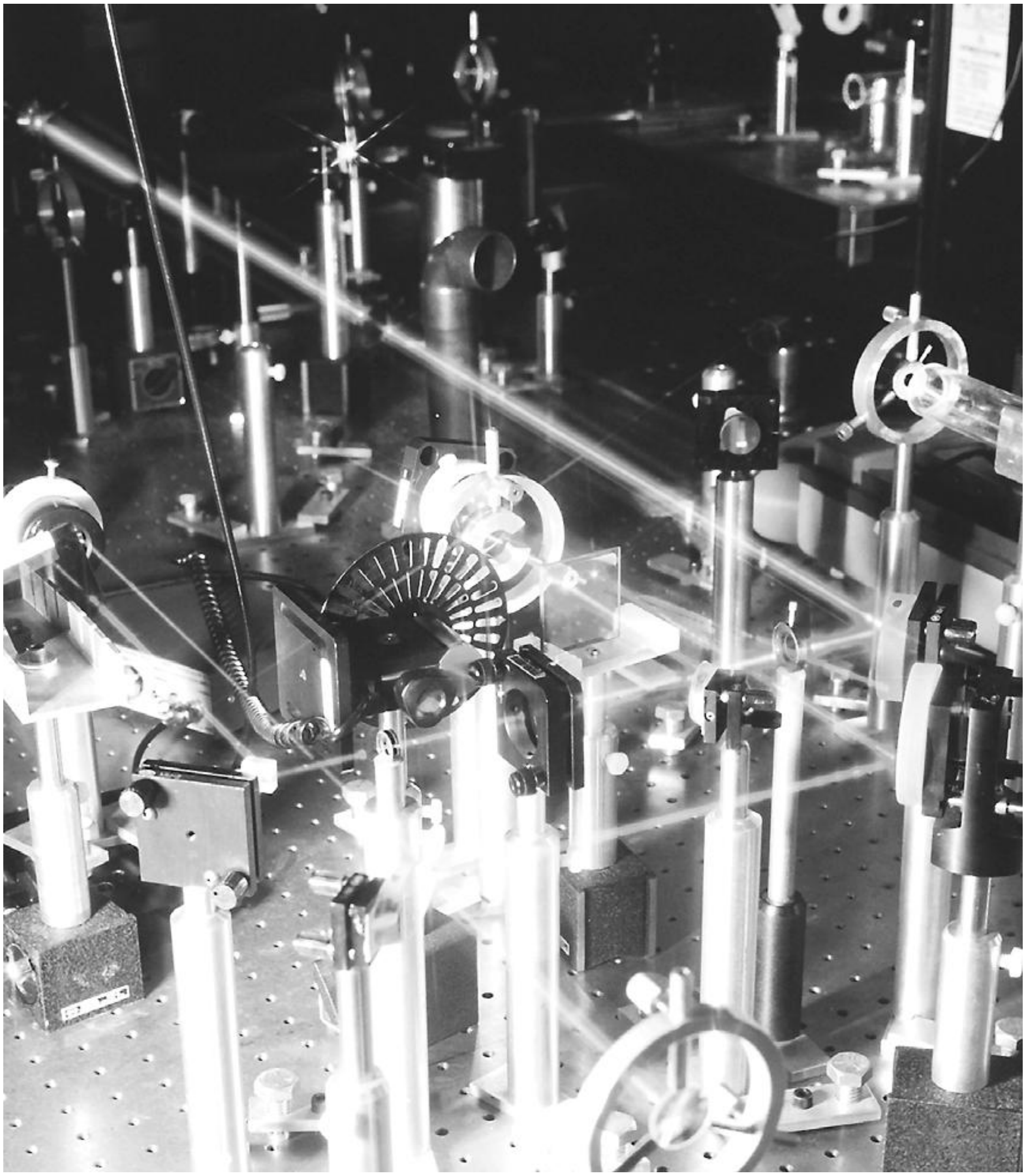
Amazingly, only about 10% of the energy output from a typical incandescent light bulb is in the form of visible light; the rest comes through the infrared region of the spectrum, producing heat rather than light that people can use. The visible light tends to be in the red and yellow end of the spectrum—closer to infrared—but a blue-tinted bulb helps to absorb some of the red and yellow, providing a color balance. This, however, only further diminishes the total light output, and, hence, in many applications today, fluorescent light takes the place of incandescent light.

Lasers

A laser is an extremely focused, extremely narrow, and extremely powerful beam of light. Actually, the term laser is an acronym, standing for *L*ight *A*mplification by *S*imulated *E*mission of *R*adiation. Simulated emission involves bringing a large number of atoms into what is called an "excited state." Generally, most atoms are in a ground state, and are less active in their movements, but the energy source that activates a laser brings about population inversion, a reversal of the ratios, such that the majority of atoms within the active medium are in an excited rather than a ground state. To visualize this, picture a popcorn popper, with the excited atoms being the popping kernels, and the ground-state atoms the ones remaining unpopped. As the atoms become excited, and the excited atoms outnumber the ground ones, they start to cause a multiplication of the resident photons. This is simulated emission.

A laser consists of three components: an optical cavity, an energy source, and an active medium. To continue the popcorn analogy, the "popper" itself—the chamber which holds the laser—is the optical cavity, which, in the case of a laser, involves two mirrors facing one another. One of these mirrors fully reflects light, whereas the other is a partly reflecting mirror. The light not reflected by the second mirror escapes as a highly focused beam. As with the popcorn popper, the power source involves electricity, and the active medium is analogous to the oil in a conventional popper.

TYPES OF LASERS. There are four types of lasers: solid-state, semiconductor, gas, and dye. Solid-state lasers are generally very large and extremely powerful. Having a crystal or glass housing, they have been implemented in nuclear energy research, and in various areas of industry. Whereas solid-state lasers can be as long as a city block, semiconductor lasers can be smaller than



THE BEAM OF A MODE LOCKED, FREQUENCY DOUBLED ND YLF LASER IS REFLECTED OFF MIRRORS AND THROUGH FILTERS AT COLORADO STATE UNIVERSITY. *(Photograph by Chris/RogersBikderberg. The Stock Market. Reproduced by permission.)*

the head of a pin. Semiconductor lasers (involving materials such as arsenic that conduct electricity, but do not do so as efficiently as the metals typically used as conductors) are applied for the intricate work of making compact discs and computer microchips.

Gas lasers contain carbon dioxide or other gases, activated by electricity in much the same way the gas in a neon sign is activated. Among their applications are eye surgery, printing, and scanning. Finally, dye lasers, as their name suggests, use different colored dyes. (Laser light itself, unlike ordinary light, is monochromatic.) Dye lasers can be used for medical research, or for fun—as in the case of laser light shows held at parks in the summertime.

LASER APPLICATIONS. Laser beams have a number of other useful functions, for instance, the production of compact discs (CDs). Lasers etch information onto a surface, and because of the light beam's qualities, can record far more information in much less space

KEY TERMS

APERTURE: An opening.

DIFFRACTION: The bending of waves around obstacles, or the spreading of waves by passing them through an aperture.

DIFFUSION: A process by which the concentration or density of something is decreased.

ELECTROMAGNETIC SPECTRUM: The complete range of electromagnetic waves on a continuous distribution from a very low range of frequencies and energy levels, with a correspondingly long wavelength, to a very high range of frequencies and energy levels, with a correspondingly short wavelength. Included on the electromagnetic spectrum are long-wave and short-wave radio; microwaves; infrared, visible, and ultraviolet light; x rays, and gamma rays.

ELECTROMAGNETIC WAVE: A transverse wave with electric and magnetic fields that emanate from it. The directions of these fields are perpendicular to one another, and both are perpendicular to the line of propagation for the wave itself.

FREQUENCY: The number of waves passing through a given point during the interval of one second. The higher the frequency, the shorter the wavelength.

HERTZ: A unit for measuring frequency, named after nineteenth—century German physicist Heinrich Rudolf Hertz (1857-1894).

INCIDENCE: Contact between a ray—for example, a light ray—and a surface. Types of incidence include reflection and refraction.

MEDIUM: A substance through which light travels, such as air, water, or glass. Because light moves by radiation, it does not require a medium, and, in fact, movement through a medium slows the speed of light somewhat.

OPTICS: An area of physics that studies the production and propagation of light.

PHOTOELECTRIC EFFECT: The phenomenon whereby certain metals acquire an electrical potential when exposed to light.

PHOTON: A particle of electromagnetic radiation—for example, light—carrying a specific amount of energy, measured in electron volts (eV).

PRISM: A three-dimensional glass shape used for the diffusion of light rays.

PROPAGATION: The act or state of traveling from one place to another.

RADIATION: The transfer of energy by means of electromagnetic waves, which require no physical medium (for example, water or air) for the transfer. Earth receives the Sun's energy (including its light), via the electromagnetic spectrum, by means of radiation.

RAY: In geometry, a ray is that part of a straight line that extends in one direction only. The term "ray" is used to describe the directed line made by light as it moves through space.

REFLECTION: A type of incidence whereby a light ray is returned toward its source rather than being absorbed at the interface.

REFRACTION: The bending of a light ray that occurs when it passes through a dense medium, such as water or glass.

KEY TERMS CONTINUED

SPECTRUM: The continuous distribution of properties in an ordered arrangement across an unbroken range. Examples of spectra (the plural of "spectrum") include the colors of visible light, or the electromagnetic spectrum of which visible light is a part.

TRANSVERSE WAVE: A wave in which the vibration or motion is perpendicular to the direction in which the wave is moving.

VACUUM: An area of space devoid of matter, including air.

WAVELENGTH: The distance between a crest and the adjacent crest, or the trough and an adjacent trough, of a wave. The shorter the wavelength, the higher the frequency.

than the old-fashioned ways of producing phonograph records.

Lasers used in the production of CD-ROM (Read-Only Memory) disks are able to condense huge amounts of information—a set of encyclopedias or the New York metropolitan phone book—onto a disk one can hold in the palm of one's hand. Laser etching is also used to create digital videodiscs (DVDs) and holograms. Another way that lasers affect everyday life is in the field of fiber optics, which uses pulses of laser light to send information on glass strands.

Before the advent of fiber-optic communications, telephone calls were relayed on thick bundles of copper wire; with the appearance of this new technology, a glass wire no thicker than a human hair now carries thousands of conversations. Lasers are also used in scanners, such as the price-code checkers at supermarkets and various kinds of tags that prevent thefts of books from libraries or clothing items from stores. In an industrial setting, heating lasers can drill through solid metal, or in an operating room, lasers can remove gallstones or cataracts. Lasers are also used for guiding missiles, and to help building contractors ensure that walls and floors and ceilings are in proper alignment.

WHERE TO LEARN MORE

Burton, Jane and Kim Taylor. *The Nature and Science of Colors.* Milwaukee, WI: Gareth Stevens Publishing, 1998.

Glover, David. *Color and Light.* New York: Dorling Kindersley Publishing, 2001.

Kalman, Bobbie and April Fast. *Cosmic Light Shows.* New York: Crabtree Publishing, 1999.

Kurtus, Ron. "*Visible Light*" (Web site). <http://www.school-for-champions.com/science/light.html> (May 2, 2001).

"*Light Waves and Color.*" *The Physics Classroom* (Web site). <http://www.glenbrook.k12.il.us/gbssci/phys/Class/light/lighttoc.html> (May 2, 2001).

Miller-Schroeder, Patricia. *The Science and Light of Color.* Milwaukee, WI: Gareth Stevens Publishing, 2000.

Nassau, Kurt. *Experimenting with Color.* New York: F. Watts, 1997.

Riley, Peter D. *Light and Color.* New York: F. Watts, 1999.

Taylor, Helen Suzanne. *A Rainbow Is a Circle: And Other Facts About Color.* Brookfield, CT: Copper Beech Books, 1999.

"*Visible Light Waves*" *NASA: National Aeronautics and Space Administration* (Web site). <http://imagers.gsfc.nasa.gov/ems/visible.html> (May 2, 2001).

LUMINESCENCE

CONCEPT

Luminescence is the generation of light without heat. There are two principal varieties of luminescence, fluorescence and phosphorescence, distinguished by the delay in reaction to external electromagnetic radiation. The ancients observed phosphorescence in the form of a glow emitted by the oceans at night, and confused this phenomenon with the burning of the chemical phosphor, but, in fact, phosphorescence has nothing at all to do with burning. Likewise, fluorescence, as applied today in fluorescent lighting, involves no heat—thus creating a form of lighting more efficient than that which comes from incandescent bulbs.

HOW IT WORKS

Radiation

Elsewhere in this volume, the term "radiation" has been used to describe the transfer of energy in the form of heat. In fact, radiation can also be described, in a more general sense, as anything that travels in a stream, whether that stream be composed of subatomic particles or electromagnetic waves.

Many people think of radiation purely in terms of the harmful effects produced by radioactive materials—those subject to a form of decay brought about by the emission of high-energy particles or radiation, including alpha particles, beta particles, or gamma rays. These high-energy forms of radiation are called ionizing radiation, because they are capable of literally ripping through some types of atoms, removing electrons and leaving behind a string of ions.

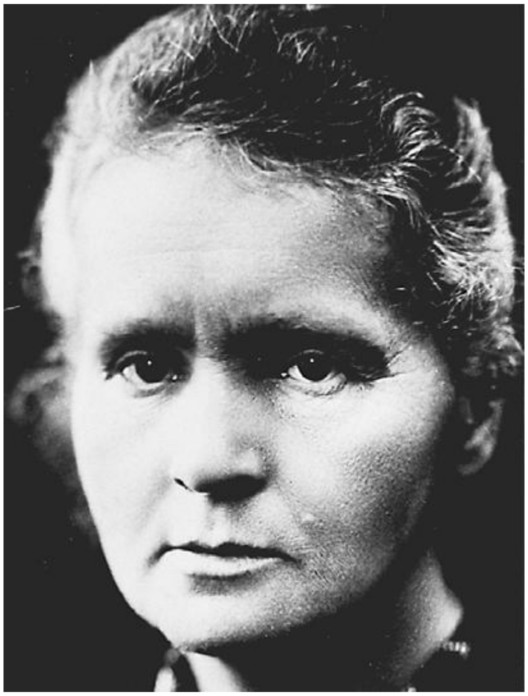
Ionizing radiation can indeed cause a great deal of damage to matter—including the matter in a human body. Even radiation produced by parts of the electromagnetic spectrum possessing far less energy than gamma rays can be detrimental, as will be discussed below. In general, however, there is nothing inherently dangerous about radiation: indeed, without the radiation transmitted to Earth via the Sun's electromagnetic spectrum, life simply could not exist.

The Electromagnetic Spectrum

The electromagnetic spectrum is the complete range of electromagnetic waves on a continuous distribution from those with very low frequencies and energy levels, along with correspondingly long wavelengths, to those with very high frequencies and energy levels, with correspondingly short wavelengths.

An electromagnetic wave is transverse, meaning that even as it moves forward, it oscillates in a direction perpendicular to the line of propagation. An electromagnetic wave can thus be defined as a transverse wave with mutually perpendicular electrical and magnetic fields that emanate from it. Though their shape is akin to that of waves on the ocean, electromagnetic waves travel much, much faster than any waves that human eyes can see. Their speed of propagation in a vacuum is equal to that of light: 186,000 mi (299,339 km) per second.

PARTS OF THE ELECTROMAGNETIC SPECTRUM. Included on the electromagnetic spectrum are radio waves and microwaves; infrared, visible, and ultraviolet light; x rays, and gamma rays. Though each occu-



MODERN UNDERSTANDING OF LUMINESCENCE OWES MUCH TO MARIE CURIE.

pies a definite place on the spectrum, the divisions between them are not firm: as befits the nature of a spectrum, one simply "blurs" into another.

Though the Sun sends its energy to Earth in the form of light and heat from the electromagnetic spectrum, not everything within the spectrum is either "bright." The "bright" area of the spectrum—that is, the band of visible light—is incredibly small, equal to about 3.2 parts in 100 billion. This is like comparing a distance of 16 ft (4.8 m) to the distance between Earth and the Sun: 93 million miles ($1.497 \cdot 10^9$ km).

When electromagnetic waves of almost any frequency are asborbed in matter, their energy can be converted to heat. Whether or not this happens depends on the absorption mechanism. However, the realm of "heat" as it is most experienced in daily life is much smaller, encompassing infrared, visible, and ultraviolet light. Below this frequency range are various types of radio waves, and above it are ultra high-energy x rays and gamma rays. Some of the heat experienced in a nuclear explosion comes from the absorption of gamma rays emitted in the nuclear reaction.

FREQUENCY AND WAVELENGTH RANGE. There is nothing arbitrary about the order in which the different types of electromagnetic waves are listed above: this is their order in terms of frequency (measured in Hertz, or Hz) and energy levels, which are directly related. This ordering also represents the reverse order (that is, from longer to shorter) for wavelength, which is inversely related to frequency.

Extremely low-energy, long-wavelength radio waves have frequencies of around 10^2 Hz, while the highest-energy, shortest-wavelength gamma rays can have frequencies of up to 10^{25} Hz. This means that these gamma rays are oscillating at the rate of 10 trillion trillion times a second!

The wavelengths of very low-energy, low-frequency radio waves can be extremely long: 10^8 centimeters, equal to 1 million meters or about 621 miles. Precisely because these wavelengths are so very long, they are hard to apply for any practical use: ordinary radio waves of the kind used for actual radio broadcasts are closer to 10^5 cm (about 328 ft).

At the opposite end of the spectrum are gamma rays with wavelengths of less than 10^{-15} centimeters—in other words, a decimal point followed by 14 zeroes and a 1. There is literally nothing in the observable world that can be compared to this figure, equal to one-trillionth of a centimeter. Even the angstrom—a unit so small it is used to measure the diameter of an atom—is 10 million times as large.

EMISSION AND ABSORPTION

The electromagnetic spectrum is not the only spectrum: physicists, as well as people who are not scientifically trained, often speak of the color spectrum for visible light. The reader is encouraged to study the essays on both subjects to gain a greater understanding of each. In the present context, however, two other types of spectra (the plural of "spectrum") are of interest: emission and absorption spectra.

Emission occurs when internal energy from one system is transformed into energy that is carried away from that system by electromagnetic radiation. An emission spectrum for any given system shows the range of electromagnetic radiation it emits. When an atom has energy transferred to it, either by collisions or as a result of exposure to radiation, it is said to be experiencing excitation, or to be "excited." Excited atoms

COMPARED TO STANDARD INCANDESCENT LIGHT BULBS (SHOWN ON THE LEFT), NEWER FLUORESCENT BULBS (SHOWN ON THE RIGHT) USE FAR LESS ELECTRICITY AND LAST MUCH LONGER. *(Photograph by Roger Ressmeyer/Corbis. Reproduced by permission.)*

will emit light of a given frequency as they relax back to their normal state. Atoms of neon, for instance, can be excited in such a way that they emit light at wavelengths corresponding to the color red, a property that finds application in neon signs.

As its name suggests, absorption has a reciprocal relationship with emission: it is the result of any process wherein the energy transmitted to a system via electromagnetic radiation is added to the internal energy of that system. Each material has a unique absorption spectrum, which makes it possible to identify that material using a device called a spectrometer. In the phenomenon of luminescence, certain materials absorb electromagnetic radiation and proceed to emit that radiation in ways that distinguish the materials as either fluorescent or phosphorescent.

REAL-LIFE APPLICATIONS

EARLY OBSERVATIONS OF LUMINESCENCE

For the most part, prior to the nineteenth century, scientists had little concept of light without heat. Even what premodern observers called "phosphorescence" was not phosphorescence as the term is used in modern science. Instead, the word was used to describe light given off in a fiery reaction that occurs when the element phosphorus is exposed to air.

There are, however, examples of luminescence in nature that had been observed from ancient times onward—for instance, the phosphorescent glow of the ocean, visible at night under certain conditions. At one time this, too, was mistakenly associated with phosphorus, which was supposedly burning in the water. In fact, the ocean's phosphorescence comes neither from phosphorus nor water, but from living creatures called dinoflagellates. This is an example of a phenomenon known as bioluminescence—fireflies are another example—discussed below.

Modern understanding of luminescence owes much to Polish-French physicist and chemist Marie Curie (1867-1934). Operating in fields that had been dominated by men since the birth of the physical sciences, Curie distinguished herself with a number of achievements, becoming the first scientist in history to receive two Nobel prizes (physics in 1903 and chemistry in 1911). While working on her doctoral thesis,

Like many marine creatures, jellyfish produce their own light through phosphorescence. *(Photograph by Mark A. Johnson. The Stock Market. Reproduced by permission.)*

Curie noted that calcium fluoride glows when exposed to a radioactive material known as radium.

Curie—who also coined the term "radioactivity"—helped spark a revolution in science and technology. As a result of her work and the discoveries of others who followed, interest in luminescence and luminescent devices grew. Today, luminescence is applied in a number of devices around the household, most notably in television screens and fluorescent lights.

Fluorescence

As indicated in the introduction to this essay, the difference between the two principal types of luminescence relates to the timing of their reactions to electromagnetic radiation. Fluorescence is a type of luminescence whereby a substance absorbs radiation and almost instantly begins to re-emit the radiation. (Actually, the delay is 10^{-6} seconds, or a millionth of a second.) Fluorescent luminescence stops within 10^{-5} seconds after the energy source is removed; thus, it comes to an end almost as quickly as it begins.

Usually, the wavelength of the re-emitted radiation is longer than the wavelength of the radiation the substance absorbed. British mathematician and physicist George Gabriel Stokes (1819-1903), who coined the term "fluorescence," first discovered this difference in wavelength. However, in a special type of fluorescence known as resonance fluorescence, the wavelengths are the same. Applications of resonance include its use in analyzing the flow of gases in a wind tunnel.

BLACK LIGHTS AND FLUORESCENCE. A "black light," so called because it emits an eerie bluish-purple glow, is actually an ultraviolet lamp, and it brings out vibrant colors in fluorescent materials. For this reason, it is useful in detecting art forgeries: newer paint tends to fluoresce when exposed to ultraviolet light, whereas older paint does not. Thus, if a forger is trying to pass off a painting as the work of an Old Master, the ultraviolet lamp will prove whether the artwork is genuine or not.

Another example of ultraviolet light and fluorescent materials is the "black-light" poster, commonly associated with the psychedelic rock music of the late 1960s and early 1970s. Under ordinary visible light, a black-light poster does not look particularly remarkable, but when exposed to ultraviolet light in an environment in which visible light rays are not propagated (that is, a darkened room), it presents a dazzling array of colors. Yet, because they are fluorescent, the moment the black light is turned off, the colors of the poster cease to glow. Thus, the poster, like the light itself, can be turned "on" and "off," simply by activating or deactivating the ultraviolet lamp.

RUBIES AND LASERS. Fluorescence has applications far beyond catching art forgers or enhancing the experience of hearing a Jimi Hendrix album. In 1960, American physicist Theodore Harold Maiman developed the first laser using a ruby, a gem that exhibits fluorescent characteristics. A laser is a very narrow, highly focused, and extremely powerful beam of light used for everything from etching data on a surface to performing eye surgery.

Crystalline in structure, a ruby is a solid that includes the element chromium, which gives the gem its characteristic reddish color. A ruby exposed to blue light will absorb the radiation and go into an excited state. After losing some of the absorbed energy to internal vibrations, the ruby passes through a state known as metastable before dropping to what is known as the ground state, the lowest energy level for an atom or molecule. At that point, it begins emitting radiation on the red end of the spectrum.

The ratio between the intensity of a ruby's emitted fluorescence and that of its absorbed radiation is very high, and, thus, a ruby is described as having a high level of fluorescent efficiency. This made it an ideal material for Maiman's purposes. In building his laser, he used a ruby cylinder which emitted radiation that was both coherent, or all in a single direction, and monochromatic, or all of a single wavelength. The laser beam, as Maiman discovered, could travel for thousands of miles with very little dispersion—and its intensity could be concentrated on a small, highly energized pinpoint of space.

FLUORESCENT LIGHTS. By far the most common application of fluorescence in daily life is in the fluorescent light bulb, of which there are more than 1.5 billion operating in the United States. Fluorescent light stands in contrast to incandescent, or heat-producing, electrical light. First developed successfully by Thomas Edison (1847-1931) in 1879, the incandescent lamp quite literally transformed human life, making possible a degree of activity after dark that would have been impractical in the age of gas lamps. Yet, incandescent lighting is highly inefficient compared to fluorescent light: in an incandescent bulb, fully 90% of the energy output is wasted on heat, which comes through the infrared region.

A fluorescent bulb consuming the same amount of power as an incandescent bulb will produce three to five times more light, and it does this by using a phosphor, a chemical that glows when exposed to electromagnetic energy. (The term "phosphor" should not be confused with phosphorescence: phosphors are used in both fluorescent and phosphorescent applications.) The phosphor, which coats the inside surface of a fluorescent lamp, absorbs ultraviolet light emitted by excited mercury atoms. It then re-emits the ultraviolet light, but at longer wavelengths—as visible light. Thanks to the phosphor, a fluorescent lamp gives off much more light than an incandescent one, and does so without producing heat.

PHOSPHORESCENCE. In contrast to the nearly instantaneous "on-off" of fluorescence, phosphorescence involves a delayed emission of radiation following absorption. The delay may take as much as several minutes, but phosphorescence continues to appear after the energy source has been removed. The hands and numbers of a watch that glows in the dark, as well as any number of other items, are coated with phosphorescent materials.

Television tubes also use phosphorescence. The tube itself is coated with phosphor, and a narrow beam of electrons causes excitation in a small portion of the phosphor. The phosphor then emits red, green, or blue light—the primary colors of light—and continues to do so even after the electron beam has moved on to another region of phosphor on the tube. As it scans across the tube, the electron beam is turned rapidly on and off, creating an image made up of thousands of glowing, colored dots.

PHOSPHORESCENCE IN SEA CREATURES. As noted above, one of the first examples of luminescence ever observed was the phosphorescent effect sometimes visible on the surface of the ocean at night—an effect that scientists now know is caused by materials in the bodies of organisms known as dinoflagellates. Inside the body of a dinoflagellate are the substances luciferase and luciferin, which chemically react with oxygen in the air above the water to produce light with minimal heat levels. Though

KEY TERMS

ABSORPTION: The result of any process wherein the energy transmitted to a system via electromagnetic radiation is added to the internal energy of that system. Each material has a unique absorption spectrum, which makes it possible to identify that material using a device called a spectrometer. (Compare absorption to emission.)

ELECTROMAGNETIC SPECTRUM: The complete range of electromagnetic waves on a continuous distribution from a very low range of frequencies and energy levels, with a correspondingly long wavelength, to a very high range of frequencies and energy levels, with a correspondingly short wavelength. Included on the electromagnetic spectrum are long-wave and short-wave radio; microwaves; infrared, visible, and ultraviolet light; x rays, and gamma rays.

ELECTROMAGNETIC WAVE: A transverse wave with electric and magnetic fields that emanate from it. These waves are propagated by means of radiation.

EMISSION: The result of a process that occurs when internal energy from one system is transformed into energy that is carried away from it by electromagnetic radiation. An emission spectrum for any given system shows the range of electromagnetic radiation it emits. (Compare emission to absorption.)

EXCITATION: The transfer of energy to an atom, either by collisions or due to radiation.

FLUORESCENCE: A type of luminescence whereby a substance absorbs radiation and begins to re-emit the radiation 10^{-6} seconds after absorption. Usually the wavelength of emission is longer than the wavelength of the radiation the substance absorbed. Fluorescent luminescence stops within 10^{-5} seconds after the energy source is removed.

FREQUENCY: The number of waves passing through a given point during the interval of one second. The higher the frequency, the shorter the wavelength.

dinoflagellates are microscopic creatures, in large numbers they produce a visible glow.

Nor are dinoflagellates the only bioluminescent organisms in the ocean. Jellyfish, as well as various species of worms, shrimp, and squid, all produce their own light through phosphorescence. This is particularly useful for creatures living in what is known as the mesopelagic zone, a range of depth from about 650 to 3,000 ft (200-1,000 m) below the ocean surface, where little light can penetrate.

One interesting bioluminescent sea creature is the cypridina. Resembling a clam, the cypridina mixes its luciferin and luciferase with sea water to create a bright bluish glow. When dried to a powder, a dead cypridina can continue to produce light, if mixed with water. Japanese soldiers in World War II used the powder of cypridina to illuminate maps at night, providing themselves with sufficient reading light without exposing themselves to enemy fire.

Processes that Create Luminescence

The phenomenon of bioluminescence actually goes beyond the frontiers of physics, into chemistry and biology. In fact, it is a subset of chemiluminescence, or luminescence produced by chemical reactions. Chemiluminescence is, in

KEY TERMS CONTINUED

HERTZ: A unit for measuring frequency, named after nineteenth-century German physicist Heinrich Rudolf Hertz (1857-1894).

LUMINESCENCE: The generation of light without heat. There are two principal varieties of luminescence, fluorescence and phosphorescence.

PHOSPHORESCENCE: A type of luminescence involving a delayed emission of radiation following absorption. The delay may take as much as several minutes, but phosphorescence continues to appear after the energy source has been removed.

PROPAGATION: The act or state of travelling from one place to another.

RADIATION: In a general sense, radiation can refer to anything that travels in a stream, whether that stream be composed of subatomic particles or electromagnetic waves.

RADIOACTIVE: A term describing materials which are subject to a form of decay brought about by the emission of high-energy particles or radiation, including alpha particles, beta particles, or gamma rays.

SPECTRUM: The continuous distribution of properties in an ordered arrangement across an unbroken range. Examples of spectra (the plural of "spectrum") include the colors of visible light, the electromagnetic spectrum of which visible light is a part, as well as emission and absorption spectra.

TRANSVERSE WAVE: A wave in which the vibration or motion is perpendicular to the direction in which the wave is moving.

VACUUM: An area of space devoid of matter, including air.

WAVELENGTH: The distance between a crest and the adjacent crest, or the trough and an adjacent trough, of a wave. The shorter the wavelength, the higher the frequency.

turn, one of several processes that can create luminescence.

Many of the types of luminescence discussed above are described under the heading of electroluminescence, or luminescence involving electromagnetic energy. Another process is triboluminescence, in which friction creates light. Though this type of friction can produce a fire, it is not to be confused with the heat-causing friction that occurs when flint and steel are struck together.

Yet another physical process used to create luminescence is sonoluminescence, in which light is produced from the energy transmitted by sound waves. Sonoluminescence is one of the fields at the cutting edge in physics today, and research in this area reveals that extremely high levels of energy may be produced in small areas for very short periods of time.

WHERE TO LEARN MORE

Birch, Beverley. *Marie Curie: Pioneer in the Study of Radiation.* Milwaukee, WI: Gareth Stevens Children's Books, 1990.

Evans, Neville. *The Science of a Light Bulb.* Austin, TX: Raintree Steck-Vaughn Publishers, 2000.

"Luminescence." Slider.com (Web site). <http://www.slider.com/enc/32000/luminescence.html> (May 5, 2001).

"Luminescence." Xrefer (Web site). <http://www.xrefer.com/entry/642646> (May 5, 2001).

Macaulay, David. *The New Way Things Work.* Boston: Houghton Mifflin, 1998.

Pettigrew, Mark. *Radiation.* New York: Gloucester Press, 1986.

Simon, Hilda. *Living Lanterns: Luminescence in Animals.* Illustrated by the author. New York: Viking Press, 1971.

Skurzynski, Gloria. *Waves: The Electromagnetic Universe.* Washington, D.C.: National Geographic Society, 1996.

Suplee, Curt. *Everyday Science Explained.* Washington, D.C.: National Geographic Society, 1996.

"UV-Vis Luminescence Spectroscopy" (Web site). <http://www.shu.ac.uk/virtual_campus/courses/241/lumin1.html> (May 5, 2001).