SCIENTIFIC AMERICAN

JULY 1993
$3.95

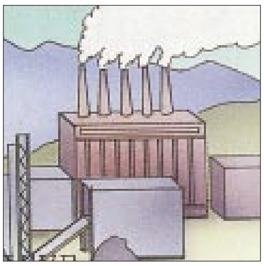
Telling reasonable risks from fools' chances.

Can sustainable development save the Amazon?

A tick in time: the most precise clocks ever.

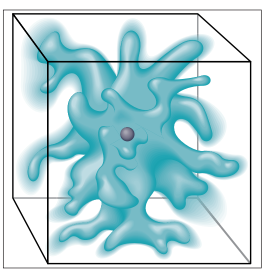
Polar dinosaurs, *which were adapted to the cold and dark, may have outlived their relatives from warmer climates.*

SCIENTIFIC AMERICAN
July 1993 Volume 269 Number 1

32 

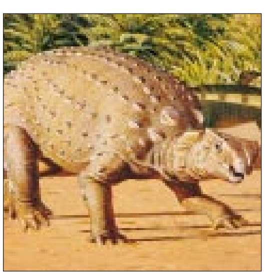
Risk Analysis and Management
M. Granger Morgan

We who live in industrial societies are justly concerned about risk. Hazards as diverse as AIDS, asbestos in schools and contamination of food and water threaten life and health. As individuals, we daily juggle the chances we take traveling, eating, smoking, drinking and encountering pathogens. Fortunately, powerful analytic techniques exist that allow policymakers to assess risk.

42 

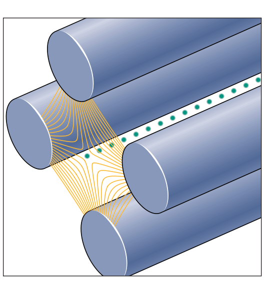
Viral Quasispecies
Manfred Eigen

The extreme mutability and adaptability of viruses wreaks havoc with the classical notion of species. But where traditional taxonomy has failed, mathematics may succeed. The author has developed a statistical classification scheme that provides insights into the evolution of the influenza virus and the age and origin of HIV, suggesting new strategies for combating viral diseases, including AIDS.

50 

Australia’s Polar Dinosaurs
Patricia Vickers-Rich and Thomas Hewitt Rich

Australia’s ability to produce varieties of animals that can be found nowhere else began at least 100 million years ago, when the continent was one with Antarctica. At that time, it hosted a population of warm-blooded dinosaurs that had evolved in the cool climate and dark winters of the polar regions. Their adaptations may have enabled them to survive longer than others in the frigid late Cretaceous period.

56 

Accurate Measurement of Time
Wayne M. Itano and Norman F. Ramsey

A spring-driven watch will get you to the church on time. A cesium clock that loses no more than a second in one million years can time the millisecond flashes of a pulsar. Yet ever more exacting scientific and navigational requirements demand even more accurate timepieces. Incredibly, such devices can be made—from isolated ions suspended in a vacuum and fountains of atoms cooled by lasers.

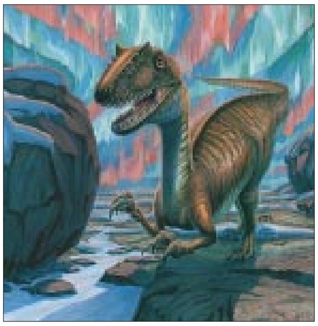
68 

Surgical Treatment of Cardiac Arrhythmias
Alden H. Harken

A number of heart attack survivors later experience an electrical short circuit that leads to an erratic, rapid heartbeat and sudden death. Because a courageous banker allowed the author and his colleagues to attempt an untried operation in 1978, the lives of most susceptible patients can now be prolonged. The pathway of the aberrant impulses is identified and surgically interrupted.

Scientific American (ISSN 0036-8733), published monthly by Scientific American, Inc., 415 Madison Avenue, New York, N.Y. 10017-1111. Copyright © 1993 by Scientific American, Inc. All rights reserved. No part of this issue may be reproduced by any mechanical, photographic or electronic process, or in the form of a phonographic recording, nor may it be stored in a retrieval system, transmitted or otherwise copied for public or private use without written permission of the publisher. Second-class postage paid at New York, N.Y., and at additional mailing offices. Authorized as second-class mail by the Post Office Department, Ottawa, Canada, and for payment of postage in cash. Canadian GST No. R 127387652. Subscription rates: one year $36 (outside U.S. and possessions add $11 per year for postage). Subscription inquiries: U.S. and Canada 800-333-1199; other 515-247-7631. Postmaster: Send address changes to Scientific American, Box 3187, Harlan, Iowa 51537. Reprints available: write Reprint Department, Scientific American, Inc., 415 Madison Avenue, New York, N.Y. 10017-1111, or fax: (212) 355-0408.

DEPARTMENTS



THE COVER painting shows *Allosaurus* hunting by the southern lights in southeastern Australia more than 100 million years ago, when the region fell within the Antarctic Circle. This specimen is one of the smallest allosaurs, and certainly the latest surviving, yet discovered. It may have owed its longevity to adaptations to cold and darkness—the very factors thought to have driven the dinosaurs to extinction some 65 million years ago (see "Australia's Polar Dinosaurs," by Patricia Vickers-Rich and Thomas Hewitt Rich, page 50).

THE ILLUSTRATIONS

Cover painting by Dimitry Schidlovsky

Page	Source
32–33	Nova Press/Sygma
34–35	Jana Brenning (*top*), Armistead Russell and Gregory McRae (*bottom*)
38–41	Johnny Johnson
43–48	Jared Schneidman/JSD
49	Jean-Pierre Prével/Sygma
50–51	Peter Trusler; for a stamp issue entitled "Australia's Dinosaur Era" by Australia Post (October 1993)
52	Johnny Johnson (*left*), Frank Coffa (*right*)
53	Patricia J. Wynne (*top*), Steve Morton, Monash University (*bottom*)
54	Peter Trusler; courtesy of *Wildlife in Gondwana,* © Reed Books International (in press)
55	Peter Menzel; courtesy of National Geographic Society (*left*), Frank Coffa (*right*)
56–57	Boris Starosta (*top*), National Institute of Standards and Technology (NIST) (*bottom*)
58	Crown copyright, Science Museum of London
59	Johnny Johnson
62	Boris Starosta (*top* and *bottom*)
63	Boris Starosta (*left*), NIST (*right*)
64	Boris Starosta
65	Johnny Johnson
68–70	Roberto Osti
72–73	Laurie Grace (*top*), Roberto Osti (*bottom*)
74	Laurie Grace
77	Michael Crawford
78–81	Ian Worpole
85	J. R. Eyerman; *LIFE Magazine,* © Time Inc.; courtesy of Henry E. Huntington Library
86	Henry E. Huntington Library (*top*), John R. Hale (*bottom right*), Henry E. Huntington Library (*bottom left*)
87–88	Patricia J. Wynne
89	Mount Wilson and Palomar Observatories
90–91	Ricardo Azoury/Black Star
92	Johnny Johnson
93	Marguerite Holloway
94	Mike Goldwater, Network/Matrix
95	Marguerite Holloway
96–99	Ricardo Azoury/Black Star
110–111	Michael Goodman

SCIENTIFIC AMERICAN®

Established 1845

EDITOR: Jonathan Piel

BOARD OF EDITORS: Alan Hall, *Executive Editor;* Michelle Press, *Managing Editor;* John Rennie, Russell Ruthen, *Associate Editors;* Timothy M. Beardsley; W. Wayt Gibbs; Marguerite Holloway; John Horgan, *Senior Writer;* Philip Morrison, *Book Editor;* Corey S. Powell; Philip E. Ross; Ricki L. Rusting; Gary Stix; Paul Wallich; Philip M. Yam

ART: Joan Starwood, *Art Director;* Edward Bell, *Art Director, Graphics Systems;* Jessie Nathans, *Associate Art Director;* Nisa Geller, *Photography Editor;* Johnny Johnson

COPY: Maria-Christina Keller, *Copy Chief;* Nancy L. Freireich; Molly K. Frances; Daniel C. Schlenoff

PRODUCTION: Richard Sasso, *Vice President, Production;* William Sherman, *Production Manager; Managers:* Carol Albert, *Print Production;* Tanya DeSilva, *Prepress;* Carol Hansen, *Composition;* Madelyn Keyes, *Systems;* Leo J. Petruzzi, *Manufacturing & Makeup;* Carl Cherebin (Ad Traffic)

CIRCULATION: Lorraine Leib Terlecki, *Circulation Director;* Joanne Guralnick, *Circulation Promotion Manager;* Rosa Davis, *Fulfillment Manager;* Katherine Robold, *Newsstand Manager*

ADVERTISING: Robert F. Gregory, *Advertising Director.* OFFICES: NEW YORK: Meryle Lowenthal, *New York Advertising Manager;* William Buchanan, *Manager, Corporate Advertising;* Peter Fisch, Randy James, Elizabeth Ryan. Michelle Larsen, *Director, New Business Development.* CHICAGO: 333 N. Michigan Avenue, Chicago, IL 60601; Patrick Bachler, *Advertising Manager.* DETROIT: 3000 Town Center, Suite 1435, Southfield, MI 48075; Edward A. Bartley, *Detroit Manager.* WEST COAST: 1554 S. Sepulveda Blvd., Suite 212, Los Angeles, CA 90025; Kate Dobson, *Advertising Manager;* Tonia Wendt. Lisa K. Carden, Lianne Bloomer, San Francisco. CANADA: Fenn Company, Inc. DALLAS: Griffith Group

MARKETING SERVICES: Laura Salant, *Marketing Director;* Diane Schube, *Promotion Manager;* Mary Sadlier, *Research Manager;* Ethel D. Little, *Advertising Coordinator*

INTERNATIONAL: EUROPE: Roy Edwards, *International Advertising Manager,* London; Vivienne Davidson, Linda Kaufman, Intermedia Ltd., Paris; Barth David Schwartz, *Director, Special Projects,* Amsterdam. SEOUL: Biscom, Inc. TOKYO: Nikkei International Ltd.

ADMINISTRATION: John J. Moeling, Jr., *Publisher;* Marie M. Beaumonte, *Business Manager*

SCIENTIFIC AMERICAN, INC.
415 Madison Avenue
New York, NY 10017
(212) 754-0550

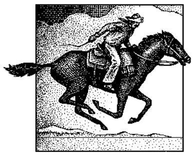
PRESIDENT AND CHIEF EXECUTIVE OFFICER: John J. Hanley

CORPORATE OFFICERS: *Executive Vice President and Chief Financial Officer,* R. Vincent Barger; *Vice Presidents:* Jonathan Piel, John J. Moeling, Jr.

CHAIRMEN OF THE BOARD: Dr. Pierre Gerckens, John J. Hanley

CHAIRMAN EMERITUS: Gerard Piel

PRINTED IN U.S.A.



LETTERS TO THE EDITORS

More Black Hole Paradoxes

I enjoyed "Black Holes and the Centrifugal Force Paradox," by Marek Artur Abramowicz [SCIENTIFIC AMERICAN, March], very much but was left somewhat puzzled. As we all know, centrifugal force is a fictitious force that appears to exist when a reference frame is rotated. For example, when a car turns a corner, unrestrained passengers tend to continue to move along a straight tangential line, in keeping with Newton's first law. Could the explanation of the paradox be made in terms of real centripetal forces?

STANLEY R. DRAKE
Shawnee, Kan.

The author writes that two astronauts inside a tube surrounding a black hole "know that the axis of the tube is circular because Bob has measured the curvature of the walls . . . using straight rulers." Is this not impossible, since linearity is defined by the path of light? Would not the ruler be unable to measure any curvature because there is no curvature along the axis of the tube?

RALF PHILIPP
Student, grade 9
Hackley School
Tarrytown, N.Y.

On page 79, the author states that "in any space-time, *with or without a gravitational field,* light always moves along geodesics, and therefore it always traces the geometry of space-time. In a space warped by a gravitational field, however, the light rays are curved and in general *do not coincide with geodesics"* (emphasis added). Is it left to the reader to choose?

GASTON FISCHER
Observatoire Cantonal
Neuchâtel, Switzerland

Every so often you publish an article that reminds me of why I subscribe. The subject matter of Abramowicz's article is fascinating, but what is particularly pleasing is that it is one of the best-written scientific articles I've ever read. Frankly, it reads like a Borges short story.

DAVID N. SCHWARTZ
London, England

Abramowicz replies:

Astronomers study rotating stars by looking at the rotating reference frame. They consider both the gravitational and centrifugal forces acting on the stellar material because the introduction of those fictitious forces makes the problem much easier. My discussion could have been in terms of free-falling frames and centripetal forces, but that would have obscured the subject.

One can tell whether two identical rulers are straight without referring to the path of light as the standard. The method is used by professional draftsmen: they simply lay the rulers beside one another. If the left and right sides of each ruler match, they are straight. Of course, the straight rulers will not appear as straight in a curved space!

Perhaps an analogy will explain why light trajectories are geodesics in four-dimensional space-time but generally not in three-dimensional space. Each great circle on a globe is a geodesic line on the two-dimensional surface, yet, being a circle, it is not a geodesic line in the three-dimensional Euclidean space in which the globe rests.

Inspecting Bridges

In "Why America's Bridges Are Crumbling" [SCIENTIFIC AMERICAN, March], Kenneth F. Dunker and Basile G. Rabbat state that "The Silver Bridge disaster [at Point Pleasant, W.Va., in 1967] happened in part because of poor inspection by local authorities." I am surprised to see that statement in *Scientific American* because there is not the slightest factual basis for it.

I was closely associated with the investigation of the collapse, beginning in January 1968 when I identified the fracture in eyebar 330 as the cause. As a metallurgical study by the National Bureau of Standards showed, the eyebar had fractured suddenly because of a stress corrosion crack less than one eighth of an inch deep that had started on the surface of the hole in the eye. The hole was almost completely filled by the pin that coupled successive links in the eyebar chain. The end of the pin and the hole in the eye were also covered by a plate that prevented visual inspection.

At the time of the collapse of the Point Pleasant bridge, an identical bridge was in service a few miles upstream. Naturally, there was great interest in determining whether its eyebars could be inspected. The best brains in the nondestructive inspection field concluded unanimously that it could not be done. Consequently, the bridge was removed.

JOHN A. BENNETT
Bethesda, Md.

Dunker and Rabbat reply:

We thank Bennett for his clarification. Ironically, lax inspection noted at the time of the Silver Bridge collapse helped to trigger a massive federal bridge inspection program, and yet state-of-the-art nondestructive testing would not have detected the hidden defect.

X-cluded from Credit

Regarding "Spot Marks the X," by John Rennie ["Science and the Citizen," SCIENTIFIC AMERICAN, April], concerning the putative role of the *Xist* gene in mammalian X chromosome inactivation, I wish to make the following clarifications. First, the human *Xist* gene was discovered during my postdoctoral fellowship at the International Institute of Genetics and Biophysics in Naples, Italy, and subsequently characterized in a collaboration with Huntington F. Willard's group and mine. Second, the murine *Xist* gene was discovered independently and reported simultaneously by Sohaila Rastan's group and mine.

ANDREA BALLABIO
Institute for Molecular Genetics
Baylor College of Medicine

ERRATA

The photograph on page 101 of "How Parasitic Wasps Find Their Hosts" [March] shows a potter wasp, which carries prey to its young, and not, as suggested, a parasitic wasp.

On page 130 of "DNA's New Twists" [March], reference is made to the "linkage of methyl groups to cysteine." The DNA base in question is actually cytosine.

Because of the volume of mail, letters to the editor cannot be acknowledged. Letters selected for publication may be edited for length and clarity.

50 AND 100 YEARS AGO

JULY 1943

"Experiments during the last eight years have led to the conclusion that atoms of gas—oxygen, hydrogen, or nitrogen—actually dissolve in the crystalline structure of some metals just as salt dissolves in water. These gas particles then 'loosen' the electrons in this structure, causing them to be emitted from the metal more readily when heat is applied. 'This explanation,' Dr. Harvey C. Rentschler recently told a meeting of the American Physical Society, 'should result in longer-lasting tubes and accomplish important savings in the size and number of electric batteries, generators, and other apparatus needed to supply the filament power.' "

"Having acquired a vitamin-B deficiency, after several weeks, does it take a person the same length of time to recover from it? Generalizations are dangerous but, by and large, the neurological and mental recovery is likely to be rapid; recovery from tissue changes, if any, probably much slower. A general idea of the former is imparted by language used in *Nutrition Reviews* with regard to deficiencies in thiamin (vitamin B1): 'There is a vast amount of evidence,' that journal states, 'that the administration of thiamin to an animal acutely deficient in thiamin causes a dramatic and prompt remission of the neurologic signs within minutes, and complete recovery within a few hours.' "

"If, as appears to be probable, vegetation exists on Mars, life has developed on two out of the three planets in our system where it has any chance to do so. With this as a guide, it appears now to be probable that the whole number of inhabited worlds within the Galaxy is considerable. To think of thousands, or even more, now appears far more reasonable than to suppose that our planet alone is the abode of life and reason. What the forms of life might be on these many worlds is a question before which even the most speculative mind may quail. Imagination, in the absence of more knowledge of the nature of life than we now possess, is unequal to the task. There is no reason, however, against supposing that, under favorable conditions, organisms may have evolved which equal or surpass man in reason and knowledge of Nature—and, let us hope, in harmony among themselves! —*Henry Norris Russell.*"

SCIENTIFIC AMERICAN

JULY 1893

"A very interesting new mammal has recently been received at the British Museum in the form of a fish-eating rat from the mountain streams of Central Peru. The animal is about the size of a common house rat, but has a flattened head, strong and numerous whisker bristles, and very small eyes and ears. The chief interest of the new form centers in the fact of its being wholly a fish-eater, and in its having in connection therewith its incisor teeth modified for catching a slippery, active prey by the development of their outer corners into long sharp points, and its intestines altered by the reduction almost to *nil* of its cæcum, an organ in vegetarian Muridæ always of great size and capacity. There is no other rodent which, as in the case of *Ichthyomys stolzmanni,* as it is proposed to name the new form, wholly lives on fish, to the exclusion of a vegetable diet."

The great Ferris wheel

"It may be of interest to amateur riflemen to know the following simple method for ascertaining the effect of gravity on a bullet: Sight the rifle upon the target, keeping the sights plumb above the center line of the bore of the rifle. Mark where the ball strikes. Then reverse the rifle, so as to have the sights exactly beneath the line of bore. In this reversed position sight it on the target as before, and mark where the bullet strikes. Divide the difference in elevation of the two bullet marks by 32 and extract the square root. This will give the time in seconds that it took the ball to travel the distance. The distance divided by this time will give the speed of the bullet per second.—*J.A.G., Grand Rapids, Mich.*"

"The Tell-el-Amarna tablets, 320 in number, were discovered by a fellah woman in 1887 among the ruins of the palace of Amenophis IV, known as Khuen-Aten, about 180 miles south of Cairo. They have been found to contain a political correspondence of the very greatest interest, dating from some 3,370 years back. Many are from Palestine, written by princes of the Amorites, Phenicians, Philistines, etc., the burden of almost all being: 'Send, I pray thee, chariots and men to keep the city of the King, my Lord.' Among the enemies against whom help is thus invoked are the *Abiri,* easily recognized as the Hebrews. The date fixes that of the Bible (I. Kings vi. 1) as accurate."

"The wonderful 'merry-go-round' designed by Engineer George W. G. Ferris, of Pittsburgh, Pa., is now completed at the Columbian Exposition in Chicago. This curious piece of mechanism forms a most remarkable and attractive object (*left*). It carries thirty-six pendulum cars, each seating forty passengers; thus one revolution of the wheel carries 1,440 people to a height of 250 feet in the air."

The Big Nada?

Inaction may stifle the UNCED agreements

The United Nations Conference on Environment and Development (UNCED) held in Rio de Janeiro last June was full of sound and fury. What exactly it signified remains to be seen. The treaties that were signed have not yet been implemented. Money that was pledged has not been forthcoming. And the group that was established to enforce Agenda 21, a 40-chapter credo for sustainable development, has not cut its teeth.

Nevertheless, many observers and former participants say it is too early to be disappointed. The Earth Summit "did not revolutionize anything. But it is a process. We are in a very large river, and its current does not change direction easily," comments Susan R. Fletcher, an expert on international environment at the Congressional Research Service, which provides data and analysis to legislators. "The major problem is that we are almost inured to rhetoric. We have heard so much about doing these things without actually doing them."

The UNCED conference, which was attended by delegates and diplomats from some 178 countries as well as by thousands of nongovernmental organizations (NGOs), resulted in the creation of a seemingly strong global political will and the endorsement of several important policy documents. Along with Agenda 21, they include the Rio Declaration (a list of environmental and development concerns that ensures national sovereignty) and a statement about protecting forests.

In addition, two conventions—one to prevent climatic change and one to conserve biodiversity—were signed by most countries. "You would still be negotiating these conventions today unless you had the driving force of UNCED," Fletcher observes. But following signatures with money and muscle is another matter. The two conventions do not become binding until they have been ratified: 50 nations must approve the climate treaty, 30 the biodiversity treaty. As of May, only 17 countries had ratified each. And if the pacts take effect but are not rigorously monitored or enforced, they will become paper tigers, like the vast majority of international environmental agreements.

Lack of enforcement could also weaken Agenda 21. Last fall the U.N. set up a 53-member Commission on Sustainable Development to oversee efforts to implement the plan. But the commission has a small staff and no legal power. It is expected to work much as the U.N. Commission on Human Rights does, by using publicity and international opinion to exert moral pressure. "There is no precedent for a group within the U.N. having the kind of clout that the Commission on Sustainable Development must have," notes Barbara J. Bramble, director of international programs at the National Wildlife Federation. "On the other hand, the U.N. is doing a lot of unprecedented things."

ALLEN TANNENBAUM *Sygma*

NONGOVERNMENTAL ORGANIZATIONS, groups representing special interests such as communities or businesses, attended the 1992 UNCED conference in Rio de Janeiro in force. Their pervasive presence at the meeting and ongoing influence on international environmental issues have been regarded as one of the most positive aspects of the Earth Summit.

At this early stage, the aspect of the fledgling commission that appears to please environmentalists and development experts the most is the inclusion of NGOs. So far some 700 organizations have asked the commission for accreditation, although NGO members such as Bramble say fewer than 100 will probably be able to maintain a presence at the U.N. A vote of one third of the members can serve to exclude an NGO—a difficulty for groups from developing countries, where some governments have tried to quell dissenting voices.

Despite potential muzzling, NGO activity is perceived as one of the Earth Summit's successful outcomes. "It is quite a victory that the rules for NGO participation are modeled on the Rio conference's rules," explains Hillary F. French, senior researcher at the Worldwatch Institute. Whether the organizations can maintain this international contact and diplomatic clout depends, in large part, on funding.

Of course, the entire enterprise of achieving the goals of the Earth Summit hinges jointly on national will and whether money can be made available to the appropriate governments, agencies and projects. Beyond the problem of getting countries to contribute 0.7 percent of their gross national product to the U.N., as Agenda 21 stipulates, lies the matter of how those funds should be used. Many developing nations "do need more money—foreign debt is a real burden," one observer remarks. "But these countries do not have a priority for sustainable development. Money going in now would go into business as usual."

Many environmentalists believe there is plenty of money around and that it is just being used incorrectly. "If you were doing a better job with what you had, you would need a lot less," notes Ian A. Bowles, legislative director of Conservation International. With regard to protecting biodiversity, for example, he argues, "we don't need a radically bigger amount of money. We just need to have it programmed right."

Funding is at the center of debates about the future of the Global Environment Facility (GEF). The GEF was established before the Earth Summit to channel funds for projects in four areas—preventing climatic change as well as protecting biodiversity, oceans and the ozone layer—that could serve as models for sustainable development. Under the joint administration of the World Bank, the U.N. Environment Program and the U.N. Development Program, the fund's pilot program is in the process of distributing $1.3 billion by December.

At that time, the future of the GEF will be reviewed. Many organizations argue that its projects are too big and uncreative and that an alternative fund should be instituted. In addition, these groups contend that the GEF's association with the World Bank ensures environmental insensitivity. The bank has been sharply criticized for the environmental damage caused by projects it has supported. "We think the GEF should be independent," says Elizabeth Barratt-Brown, a senior attorney at the Natural Resources Defense Council. "There has been a lot of greening in the language of the World Bank. But what is really changing in the funding?"

Other organizations, such as Conservation International, believe the GEF should remain affiliated with the bank and serve as a lever to influence change there. "It is an inherently long-term proposition, but all these things are continually advancing. The U.N. and the World Bank have been around forever, so it is hard to reform suddenly," Bowles points out.

One notable change since the Earth Summit involves the U.S.'s position on the environment. During his tenure, President George Bush refused to sign the convention on biodiversity in Rio and was unwilling to meet the goals of the climate convention to lower carbon emissions, which contribute to global warming, to 1990 levels by the year 2000. During his Earth Day address this past spring, however, President Bill Clinton emphasized his willingness to meet the targets for carbon emissions.

Clinton also announced his intention to sign the convention on biodiversity, with the addition of what is called an interpretive statement. Because worries about U.S. intellectual property rights had deterred Bush from supporting the treaty, the new administration sought to confront the issue by clarifying aspects of the convention—a procedure that is common to many treaties. The interpretive statement allows "the U.S. to get a common statement that both businesses and environmentalists could live with," Bowles notes. "Ratification depends on getting the wording right in the statement."

Such changes in national policy seem to be rare. Some communities and countries, notably the Philippines, have tried to establish local and national sustainability. But in general, "we have seen a return to business as usual around the world," says Jacob Scherr, a senior staff attorney at the Natural Resources Defense Council. "These international treatises demand an enormous amount of attention and energy and should not be a diversion from needed efforts on the ground." —*Marguerite Holloway*

Moonball

Astronomers beat a path to high resolution

Harold A. McAlister of Georgia State University is an astronomer, but he frequently finds himself talking baseball when he describes his work. "If you built a stadium on the moon, you couldn't even see it from the earth through the best optical telescopes," he begins. But McAlister is championing a technique called optical interferometry that would allow earthbound fans to watch the game. "With our array, you could see who's pitching."

By bringing together beams of starlight captured by two or more widely separated telescopes, McAlister and his colleagues can achieve the equivalent resolving power of a single instrument as large as the distance between the telescopes. When the beams are combined, the light waves interfere with one another. Where the peak of one light wave meets the peak of another, they reinforce each other; where the peak of one wave meets the trough of another, they cancel out.

An electronic detector records the resulting pattern of dark and light areas, or interference fringes, which can then be analyzed by computer to extract detailed information about the object being observed. If at least three telescopes are used, the fringes can be rendered into images hundreds of times crisper than even those from the orbiting *Hubble Space Telescope*—at perhaps one hundredth the cost.

Many of the most impressive results reported so far have come from the Mark III Optical Interferometer on Mount Wilson in California, which has been operating since 1986. The Mark III consists of two mobile light collectors that can be placed as far as 31 meters apart. The longer the distance between the individual telescopes, the greater the instrument's resolving power. At full extension, the Mark III can pick out details as small as two thousandths of an arc second, about 100,000 times better than the human eye can.

The Mark III can measure the outlines of astronomical objects, but, alas, it cannot make true images. Nevertheless, it has proved the importance of the concept. Last year Nicholas M. Elias and his colleagues at the U.S. Naval Observatory made a stunning measurement of a shell of gas blasting away from Nova Cygni 1992, a brilliant thermonuclear detonation that occurred on the surface of a collapsed white dwarf star. Elias found that 10 days after the explosion the shell stretched to 3.8 thousandths of an arc second (the full moon, in comparison, is 1,800 arc seconds in diameter). Observations of Nova Cygni's spectrum revealed the velocity of the fleeing gas. Combining those data with the Mark III measurement enabled Elias's group to determine that the nova is about 9,500 light-years from the earth.

Other findings from the Mark III have illuminated the shape and structure of stars. Stars are so distant in comparison to their diameters that astronomers have always considered them as unresolvable point sources of light. But optical interferometers can resolve the disks of many stars and reveal features on their surfaces. Michael Shao of the Jet Propulsion Laboratory in Pasadena, Calif., estimates that the Mark III has already resolved about 200 stellar disks.

One surprising result of observations done on the Mark III and elsewhere is the discovery that "stars are not round," says Richard S. Simon of the Naval Research Laboratory. Many red giant stars, including the bright star Betelgeuse and the well-known variable star Mira, exhibit peculiar egglike shapes, presumably because of the huge convection currents roiling their filmy outer layers. A team led by Simon has also reported detecting a huge cocoon of hydrogen gas surrounding the hot, highly active blue star Gamma Cassiopeia. Related work has revealed clouds of titanium oxide billowing off red giants' distended surfaces.

Future optical interferometers promise to push the technology and yield even grander results. A group at the University of Sydney led by John Davis is busily completing a 640-meter-long optical array that will be able to measure stellar diameters as small as 50 millionths of an arc second (some 40 times better than the Mark III and about 1,000 times smaller than the finest details visible to the *Hubble Space Telescope*).

"One of our key goals will be measuring the pulsations of Cepheids," Davis relates. Cepheids are a class of pulsating stars whose regular variations in brightness have been used by cosmologists to establish the distances to remote galaxies. Davis hopes to correlate direct measurements of Cepheid pulsations with spectroscopic observations of how fast their surfaces rise and fall. In much the way that Elias inferred the distance to Nova Cygni, Davis and his colleagues will use their information to derive distances to Cepheids—and thus help calibrate the size of the universe.

The most exciting results will come from interferometers that link multiple telescopes and thereby deliver the long-elusive goal of creating true images. Craig Mackay of the University of Cambridge expects that the university's 100-meter-long interferometer, known as COAST, will begin producing such images later this year. The instrument initially will target familiar objects such as red giants and tightly bound binary stars "to make sure we're not totally off base," Mackay says cheerily. Then he hopes to train COAST on T Tauri stars, stellar newborns still enshrouded with disks of gas and dust. Resolving details around T Tauris will vastly increase astronomers' understanding of the process by which stars and planets form.

The Naval Observatory and Naval Research Laboratory are constructing a similar but larger imaging optical interferometer, the Big Optical Array, which will begin operation sometime in the fall. McAlister's planned Center for High Angular Resolution Astronomy array will incorporate seven large, 100-centi-

Banzai!

Generally, old satellites don't die; they just fade away. Yet there are exceptions. This past spring the Japanese Institute of Space and Astronautical Science (ISAS) decided to send its *Hitin* satellite into oblivion not with a whimper but a bang. Rather than flinging the aging spacecraft into the nether reaches of the galaxy, ISAS piloted it straight into the moon. On April 10, when the 315-pound probe crashed at roughly 5,600 miles per hour, it exploded in a bright flash, throwing up dust and digging out a crater that astronomers hope will serve as a new benchmark for planetary science.

Håkan Svedhem, a physicist with the European Space Agency, heard rumors of ISAS's plans two weeks before the execution date and scrambled to persuade astronomers to train their telescopes on the moon that night. "It was a great opportunity to observe from the ground a really giant impact as it happens. This has not been done before," Svedhem says.

Three observatories around the world signed on. But as the kamikaze satellite plunged toward its fiery demise, the telescope in Irkutsk was jammed up with technical difficulties, and another in Indonesia was rained out. The last hope was David Allen, an astronomer at the Anglo-Australian Observatory who has a reputation for making difficult observations. "If anybody could get this shot, David could," says Alistair Glasse of the Royal Observatory in Edinburgh. But because of miscommunication about the time of impact, Allen was unwittingly running behind schedule.

Just seconds before the collision, Allen got the cameras rolling on the observatory's infrared imaging spectrometer and recorded half a dozen frames as the flash lit up the lunar night. The intensity of the burst and the apparent lack of a sizable dust cloud make Glasse suspect that *Hitin* hit solid rock, converting nearly all its kinetic energy to heat and light. Svedhem points out, however, that because ground zero lay about 10 kilometers within the Cimmerian side of the terminator between night and day, a large dust cloud could easily have been cloaked in darkness.

The deceased was born *Muses-A,* a small craft designed to help Japanese astronavigators hone their lunar swing-by skills in preparation for a joint mission with the National Aeronautics and Space Administration. Christened *Hitin* (a goddess of music) after its launch on January 24, 1990, the satellite surpassed its one-year life expectancy and after a second year in high earth orbit was sent to wheel round the moon. While it was en route, Svedhem used the instrument to collect data on cosmic dust.

Hitin's grand finale was not intended to benefit science. ISAS officials gave only vague explanations for their decision—"something about leaving fragments for their great-grandchildren to find," Svedhem reports. But the satellite may yet attain martyrdom by providing a rare controlled experiment for planetary scientists. "The correlation between the size and velocity of a meteorite and the size of the crater it creates is based on theoretical calculations and has never been verified by observations," Svedhem explains. "In this case, we had a very well defined mass and velocity. But of course we cannot see the crater yet; it is quite small."

Svedhem hopes *Hitin*'s successor will pay a visit to the grave site and send back images of the crater. Meanwhile he and Glasse will glean all they can from their pictures of the day the *Muses* died. *—W. Wayt Gibbs* meter telescopes and will be sensitive to infrared radiation—a capability that reduces atmospheric distortion and improves sensitivity to cool objects such as dust-cloaked infant stars.

Perhaps the most audacious devices on the drawing boards are the interferometry arrays proposed to be built late in this decade around two of the largest telescopes in the world: the pair of Keck telescopes on Mauna Kea in Hawaii and the European Southern Observatory's Very Large Telescope in Chile. These devices will scrutinize the disks around young stars, explore the tortured inner regions of active galaxies and search for planets orbiting other stars. Shao estimates that even the extremely ambitious Keck array will have a price tag of $40 million—a hefty sum, but only about half the cost of each of the primary Keck telescopes and a tenth the cost of a single flight of the space shuttle.

Such funds are not immediately forthcoming, however. A cloud of penury hangs over the field of optical interferometry. Part of the problem lies in skepticism within the scientific community. "Astronomers tend to be a conservative group. A lot of people consider interferometry to be black magic," McAlister sighs. Shao hopes the newest set of devices, including his current project, a test-bed for the Keck array, "will be able to convince lots of conventional astronomers that interferometry is a tool that will be useful for them."

Of course, the astronomers are not the only ones who need convincing, as McAlister knows only too well. He anxiously awaits approval of the next dollop of funds from the National Science Foundation. "It's contingent on the federal budget," he says. "That is always risky business." —*Corey S. Powell*

QED for QCD

A supercomputer backs the theory of quarks

It's a good thing machines don't get overtime. Researchers at the IBM Thomas J. Watson Research Center have recently completed a set of calculations on a supercomputer that ran continuously for approximately an entire year. More than an exercise in patience, the task may have provided the strongest confirmation yet of a current theory of elementary particles. In particular, the IBM team calculated the masses of eight hadrons, a family of particles that includes the proton and neutron, and showed that the values obtained are consistent with the masses measured in the laboratory.

The theory of quantum chromodynamics, or QCD for short, was postulated in the 1970s to describe how the fundamental particle known as the quark builds the various hadrons. Two "up" quarks and a "down" quark, for example, create a proton. A so-called chromoelectric field (based on a property of quarks called color) holds the quarks together; the chromoelectric field is carried by particles called gluons. The QCD theory was highly successful in enunciating the properties of hadrons in certain kinds of experiments and became part of the so-called Standard Model, which unites all the forces of nature except for gravity.

Although experiments can supply data for hadrons—one can simply look up the mass of the proton in reference books—a correct theory should be able to predict such information. Deriving observed values via analytic means would greatly substantiate the model. Besides giving physicists confidence they have the right ideas, such derivations suggest that quantum parameters that cannot be detected experimentally can be accurately inferred. "We would like to get hadron masses to very good accuracy to help us look at QCD," says W. Douglas Toussaint, a physicist at the University of Arizona. "It would enable us to compute properties that are useful for extracting fundamental constants, such as quark masses."

But the theory's mathematical complexity has made such predictions almost impossible. Perturbation theory, the main tool of quantum field physics, proved inappropriate for a complete description of QCD. It applied only to a limited part of the model. Instead investigators turned to numerical methods on computer, based on lattice gauge theory, a mathematical framework erected 20 years ago by Kenneth G. Wilson, now at Ohio State University. The lattice refers to the representation of space as a scaffold, on which quarks rest on connecting sites. The bonds between lattice points represent the gluons.

To secure answers representative of the real world, workers must conduct the calculations as the distance between lattice points shrinks to zero and the number of lattice points increases to infinity. In these limits, one should be able to come up with observable quantities. Indeed, researchers have used lattice QCD to explain quark confinement, which accounts for why no one can see any free quarks: it would take an infinite amount of energy to isolate a quark.

Coming up with the masses of hadrons has proved even more elusive. "The calculations require that you look at all possible different configurations of quarks, antiquarks and the chromoelectric field on the lattice," says Donald H. Weingarten, who headed the IBM team. For meaningful results, large lattices are necessary, and that entails more involved calculations—more than 100 million billion arithmetic operations.

Hence the need for a supercomputer. Weingarten and his IBM colleagues Frank Butler, Hong Chen, Jim Sexton and Alessandro Vaccarino turned to GF-11, a massively parallel computer they helped to develop for QCD calculations. The

INTERNATIONAL BUSINESS MACHINES

***QUARK CALCULATOR** Donald H. Weingarten poses with the supercomputer he and his IBM colleagues used to compute the masses of subnuclear particles, thereby providing confirmation that the theory of quantum chromodynamics is correct.*

nomenclature refers to the maximum speed of the calculations: 11 gigaflops, where a flop is a floating decimal point operation per second. To obtain the hadron masses, the researchers ran the GF-11 for more than a year at a sustained rate of between five and seven gigaflops. That is roughly equivalent to running more than 200,000 desktop computers that use 386 processors.

The result after a yearlong wait? "Statistically, the numbers for the hadron masses are completely consistent with experiment," Weingarten says. The disagreement for some values ranged from less than 1 percent to no more than 6 percent. The uncertainties are for the most part by-products of the statistical algorithm used in the computer calculations and are not reflections of QCD. In other words, QCD seems to be the right theory.

Despite their success, Weingarten's calculations rely on a simplification often made to render the mathematics doable. Called the valence approximation method, it does not fully take into account an effect that occurs in quantum systems—the spontaneous creation and destruction of particles. Quantum fluctuations can cause quark-antiquark pairs to flash into existence and thereby influence the system in some way. Rather than incorporating the process, the valence approximation assumes that such virtual pairs act mainly to reduce the strength of the existing color field. The approximation then compensates for the decrease.

Not everyone thinks the approach is completely valid. "If you leave out a significant part of the theory, you don't know what the effect will be," maintains Norman H. Christ, a physicist at Columbia University. Christ is deriving the values with the full theory, using Columbia's supercomputer, the only other machine dedicated solely to QCD reckoning. But Toussaint estimates that calculating from the full theory would require 100 times the computer power of current machines.

The approximation does not undermine Weingarten's confidence in his results. "The agreement between the valence approximation and experiment in a sense tells you that the full theory must agree with experiment," he says.

The physicists do concur on one point: the calculations support the legitimacy of employing computers in a proof. "The sort of thing we are doing represents a real qualitative change," Weingarten muses. What is new is that investigators accept an "experimental" uncertainty in studying and testing a theory in fundamental physics. "It's experimental theoretical physics." —*Philip Yam*

A Kinder War

"Harm reduction" gains ground as an approach to drug abuse

During a presidential debate last October 11, a reporter asked the candidates whether they would consider legalizing drugs in order to reduce drug-related violence and crime. Bill Clinton rejected that option. Declaring that the life of his half brother, Roger, had been saved by his arrest for possession of cocaine (Clinton himself had authorized the arrest), the president-to-be insisted law enforcement was crucial for combating drug abuse.

Clinton backed up his tough rhetoric with his first budget. It called for spending $13 billion in the next fiscal year on controlling drugs, almost $1 billion more than the Bush administration earmarked for the current year. Clinton allocated 64 percent of the funds for antismuggling programs and law enforcement (the balance is for education and treatment), only slightly less than Bush had.

Nevertheless, critics of the so-called war on drugs are hopeful that the new administration will be willing to try different tactics. "Change is in the air," remarks Arnold S. Trebach, a professor of criminal justice at American University and president of the Drug Policy Foundation, a nonprofit group in Washington, D.C., that espouses an approach to drugs called "harm reduction."

The idea behind harm reduction is that drug abuse should be viewed as, at worst, a disease requiring treatment and not an absolute evil that must be eradicated at all costs. "The essence is the acceptance of the enduring reality of drug use, the absurdity of even attempting to create a drug-free society and the need to treat drug users and abusers as basically decent human beings," Trebach says.

Support for this viewpoint is growing at the international level. In the past three years, representatives of 15 European provinces and cities, including Zurich, Amsterdam and Rome, have signed the so-called Frankfurt Resolution, which calls for easing prohibitions on marijuana, free availability of clean needles and treatment for all who seek it. Trebach says his foundation, together with the city of Baltimore, is co-sponsoring a meeting on the resolution this November. Kurt L. Schmoke, the mayor of Baltimore, who has supported harm-reduction policies, has pledged to sign the resolution at the meeting.

The harm-reduction philosophy also pervaded a conference on drugs held in the nation's capitol on May 7 and attended by specialists in drugs, among them law enforcement officials, judges, physicians and social scientists. Although the audience disagreed on how to reduce the harm caused by illegal drugs—the proposals ranged from relatively modest calls for more treatment to outright legalization of all drugs—almost all concurred that the war waged by the Reagan and Bush administrations had been an expensive failure.

Indeed, the annual federal budget for drug war activities surged from less than $2 billion in 1981 to more than $12 billion for this fiscal year. The Bush administration alone spent more than $40 billion to suppress illegal drug use over four years. More than two thirds of the funds went toward efforts to decrease smuggling and to enforce laws.

Federal and state governments also instituted more severe penalties for drug violations, including mandatory sentences for those convicted of possession or sale of drugs exceeding certain amounts. Consequently, the number of arrests and convictions for drug violations soared to record levels. Drug offenders account for roughly a third of the U.S. prison population, which reached an all-time high of 883,593 at the end of 1992.

Defenders of strict policies claim their implementation has reduced the number of people who dabble in illegal drugs, marijuana in particular. Surveys done by the National Institute on Drug Abuse show that casual drug use has fallen steadily since 1979. Critics contend the decreases resulted less from law enforcement efforts than from a growing public awareness of the adverse effects of all drugs, legal or illegal. They note that the use of tobacco and alcohol has also decreased over the past decade.

Moreover, crime and other problems related to drug abuse and trafficking continue unabated, particularly in poor urban communities. Overdoses and medical emergencies caused by cocaine, heroin and other drugs dropped in 1990 but have risen again since then. The rate at which intravenous drug users are infected with AIDS continues to grow [*see illustration on page 26*].

Through his appointments—if not his rhetoric—Clinton has set the stage for change. At the May conference in Washington, Attorney General Janet Reno said her experience as state attorney in Dade County, Florida, a major center of drug trafficking, led her to conclude that antismuggling programs were having little impact on the cost or availability of drugs. She also complained that mandatory sentences for nonviolent drug offenders had decreased the prison space available for incarcerating more dangerous criminals.

Reno urged that nonviolent drug offenders be handled with a "carrot and stick" approach, in which they can avoid prison by submitting to a treatment program and staying off drugs; urine tests would ensure compliance. Such a plan has been carried out in Dade County during the past four years—with great success, Reno said. This system has also been favored by Lee P. Brown, former commissioner of police in New York City, whom Clinton named head of the Office of National Drug Control Policy.

Some prominent jurists have proposed more radical measures. One is Whitman Knapp, a senior federal judge in New York State (famed for having led a commission that investigated police corruption in New York City two decades ago). Earlier this year Knapp announced he would refuse to consider drug cases subject to mandatory sentencing laws. He subsequently argued that Congress should repeal all federal laws banning drug sales or possession and permit states to devise alternatives to prohibition.

Opponents of such wholesale decriminalization fear any benefits would be offset by a tremendous upswing in the abuse of drugs such as cocaine and heroin. David F. Musto, a historian at Yale University, suggests in the 1987 book *Dealing with Drugs* that in 1900, before opioids were prohibited in the U.S., the rate of opioid addiction was at a level "never equaled or exceeded."

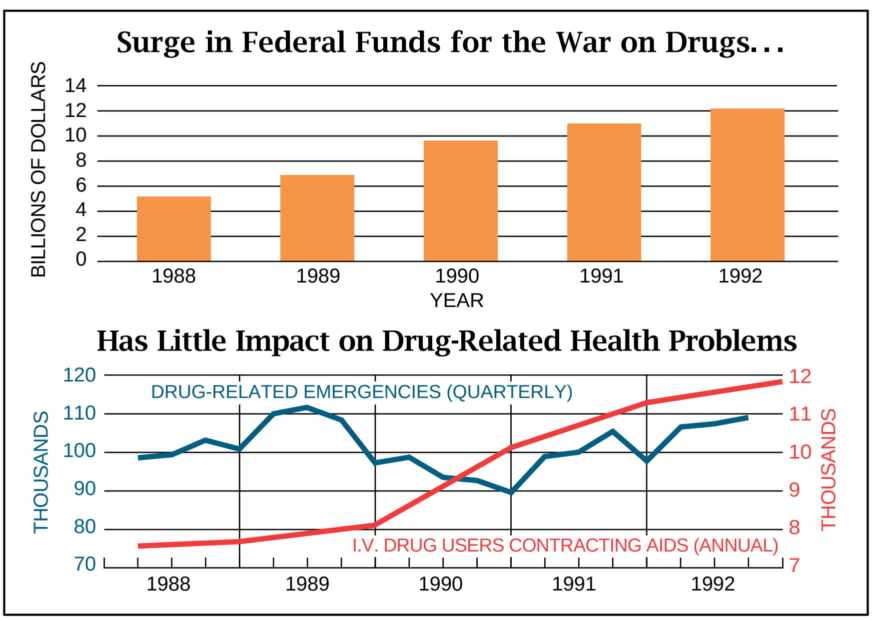
Trebach challenges this claim. He argues that estimates of the number of addicts varied wildly at the turn of the century, as do current estimates; the historical evidence can be used to buttress any conclusion. "The charge that prohibition enforced through the criminal law has succeeded in reducing the total number of addicts or the rate of opiate addiction in the United States cannot be supported by the evidence at hand," Trebach states in an upcoming book.

Some opponents of legalization highlight the benefits of Prohibition, the period from 1920 to 1933 during which alcohol was outlawed throughout the nation. Ethan A. Nadelmann, a public policy scholar at Princeton University, acknowledges that consumption of alcohol did indeed fall during Prohibition, as did public drunkenness and cirrhosis of the liver. Yet he notes that alcohol-related problems had decreased even more sharply during World War I as a result of alcohol rationing and the temperance movement. Moreover, Britain was more successful than the U.S. at reducing alcohol consumption and related health problems in the 1920s and 1930s through taxes and restrictions on hours of sale.

On the other hand, at least one recent experiment in decriminalization was a spectacular failure. Five years ago officials in Zurich designated a park in which drugs could be used without interference. Zurich recently ended the experiment after the park became a haven for dealers, prostitutes and addicts from throughout Europe.

Some experts, while ruling out wholesale decriminalization, have proposed partial measures. Mark A. R. Kleiman of the Kennedy School of Government at Harvard University suggests a policy that he calls "grudging toleration." It would allow the sale of certain psychoactive drugs through state-regulated stores but would discourage consumption through such measures as steep taxes and limits on amounts purchased. Kleiman thinks alcohol, tobacco and marijuana might all be included under this regime, but he would exclude drugs he considers too harmful, notably cocaine and heroin.

Kleiman's proposal aside, the drug reform movement has been more effective at criticizing current approaches than at suggesting specific alternatives. To redress that problem, Nadelmann has helped form the Princeton Working Group on the Future of Drug Use and Alternatives to Drug Prohibition, consisting of experts from Princeton and other institutions. A primary goal of the group, Nadelmann says, is to devise a "drug regulatory model that eliminates many of the worst consequences of drug prohibition without reproducing the unfortunate consequences of our alcohol- and tobacco-control policies." After all, Nadelmann remarks, alcohol and tobacco remain by far the most harmful of all drugs. —*John Horgan*



LAURIE GRACE

SOURCES: Centers for Disease Control (red line), *Drug Abuse Warning Network* (blue line)

Healing Hearing

Regrowing damaged ear cells might eventually cure deafness

If your taste for loud music has forced you to swap your Walkman for a hearing aid, there is now a chance that you might someday be able to switch back. The deafness caused by loud noise has traditionally been permanent. But researchers have recently found encouraging signs that humans may have at least a latent ability to regenerate damaged parts of their inner ear. Drugs that stimulate that regrowth could conceivably restore hearing.

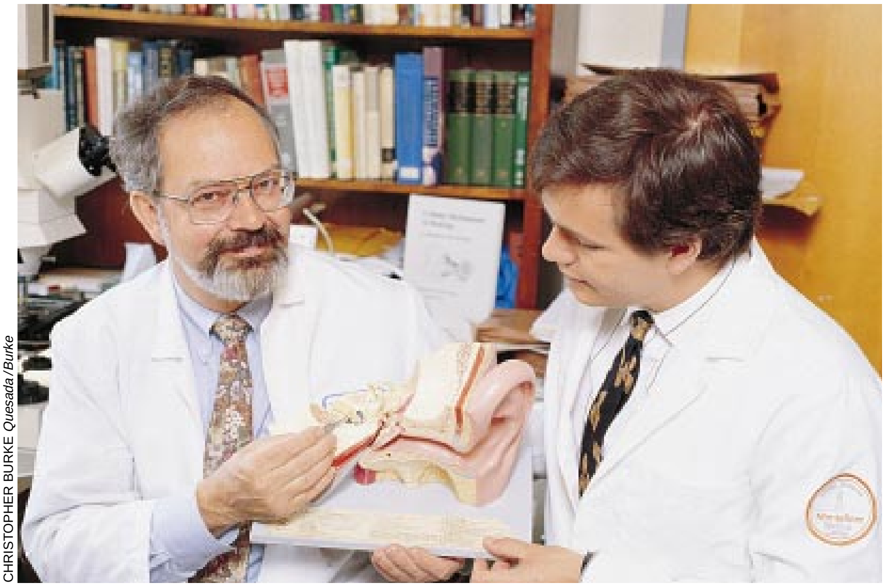
"To me, it's no longer a question of if but of when we will get regeneration in humans," predicts Jeffrey T. Corwin, a hearing investigator at the University of Virginia School of Medicine who has contributed to the new findings. Thomas R. Van De Water of Albert Einstein College of Medicine in Bronx, N.Y., agrees: "It's an exciting time. I've been working 25 years in this field, and all of a sudden it's breaking open."

The focus of their work is the cochlea, a periwinkle-shaped organ of the inner ear. When sound waves strike the eardrum, the vibrations pass into the fluid filling the cochlea and set in motion the tiny structures called hair cells that stimulate the auditory nerve. Unfortunately, hair cells are delicate: 120 decibels of Metallica (or Mahler, for that matter) jolts some hair cells so hard that they shear off from their foundation.

Antibiotics, anticancer drugs and some diseases, such as meningitis, can also kill hair cells. If too many are lost, the ear becomes irreversibly deaf.

Up to a point, hearing aids can sometimes compensate for the lost sensitivity by amplifying sounds. In recent years, cochlear implants have become available; surgeons have inserted about 8,000 of them into deaf patients in the U.S. These electronic devices allow microphones to stimulate the auditory nerve directly. They cannot restore normal hearing, however: at best, and after rehabilitative therapy, the implants permit patients to recognize sounds.

To some audiologists, a better solution would be to repair the injured cochlea. Yet in humans and other mammals, years of observation indicated that hair cells did not regenerate naturally. Other animals seemed more resilient. More than a decade ago Corwin began finding evidence that sharks and amphibians grow new hair cells throughout their lives. Birds, too, can recover from deafness. "Normally, birds don't have turnover of their hair cells," notes Douglas A. Cotanche of the Boston University School of Medicine, a pioneer of the avian work. "But if you stress them with noise or drugs, there is trauma, and new cells regenerate." For whatever reason, he says, "there was this prejudice that 'mammals are too evolved to do that.' "



CHRISTOPHER BURKE *Quesada/Burke*

***EAR CELLS CAN REGENERATE** under some conditions, according to Thomas R. Van De Water* (left) *and Hinrich Staecker* (right) *of Albert Einstein College of Medicine.*

New studies are changing that opinion. This past March in *Science,* Corwin and Andrew Forge of University College, London, and their colleagues claimed to find regenerating hair cells in the inner ear of guinea pigs. They had treated the animals repeatedly with toxic doses of the antibiotic gentamycin. Within a week, almost all the hair cells in one region were gone. Yet a month after the treatments, immature hair cells were visible. In an accompanying paper, they described what happened when they repeated the experiment in culture using layers of tissue from the inner ears of guinea pigs and of humans in their fifties. When antibiotics eliminated the hair cells, adjacent supporting cells divided and took their place. The supporting cells then differentiated as hair cells.

Skeptics could still argue that these results were not pertinent to deafness: Corwin's group had worked with hair cells from the vestibular system, which confers the sense of balance, and not from the cochlea. Van De Water and his graduate student Hinrich Staecker and a team from the University of Liège cleared that hurdle just six weeks later. Using cochlear tissue from rat pups only a few days old, they found that the hair cells did not spontaneously recover. But they also discovered that extensive regrowth could be encouraged in less than a week if they exposed the tissue to retinoic acid—a compound related to vitamin A that guides the differentiation of many cells during embryonic development. Retinoic acid derivatives are now used as wrinkle creams and treatments for severe acne and have shown some potential for fighting cancer in recent studies.

Van De Water argues that regeneration of hair cells seems to hinge on two events: the destruction of the old cells and the receipt of a chemical stimulus to spur regrowth by new ones. Mature mammals may not provide that second signal efficiently.

Although he is excited by the new results, Cotanche points out that "no one has ever really studied to see whether a regenerated hair cell is functional." The cells must not only grow back but must also connect to the nerve. Moreover, as Corwin explains, "the hair cells that are sensitive to different pitches are slightly different in their structure."

One of the "amazing" observations that have been made, Corwin says, is that the replacement cells all seem to develop the appropriate structures and orientations, apparently in response to some cues embedded in the tissue environment. "Also, all our indications are that the nerve cells can find the new sensory cells," he adds. Both he and Van De Water suspect the regenerating hair cells may be releasing chemicals that attract neuronal connections. "I've never had much concern that they were eventually going to hook up," Corwin says. "They always get hooked up in sharks and birds."

The workers caution that therapeutic regeneration of hair cells in humans will have to wait a while. Even if retinoic acid does the trick, getting it into the ear could be a problem. "We don't want people to take large amounts of vitamin A, because it can be dangerous," Van De Water warns. "We're trying to develop unique ways of delivering the drug right to the inner ear tissue using miniature osmotic pumps." He says that his laboratory is looking into the possibility of gene therapy, using viruses that could insert the genes for growth controls into cochlear cells. Cotanche wonders, too, whether immature cells might be implanted into the cochleas of deaf adults and induced to become hair cells. Another issue, Van De Water notes, is that the neurons that become disconnected from hair cells often die; he and his coworkers are trying to find out how to keep them alive until new hair cells have emerged.

"I expect we'll have some very good compounds and protocols worked out for stimulating regeneration within the next two or three years," Corwin ventures. He, Van De Water and Cotanche all agree that it will most likely be a decade before an actual therapy is ready for testing in patients. Ever the optimist, though, Corwin says, "These discoveries over the past year or so have shaved 10 years off our projections."

Ten years is a long time for some deaf people to wait. "I would love to see us put the cochlear-implant people out of business," Van De Water remarks. "I'm sure they would love it also. Right now, though, the cochlear implant is the only show in town." —*John Rennie*

Science's Unscientific Champion

When Irving S. Shapiro turned 65 in 1981, he had completed seven successful years as chairman of E. I. du Pont de Nemours & Company and earned an enviable reputation as an advocate of ethical standards in commerce. He bought a house in Florida and was ready to retire to a quiet life of golf. It was not to be. Shapiro, who was trained as a lawyer, says he "was favored with so many offers from law firms that I began to think I was selling myself short." He took a partnership at a major firm and then pulled off the most challenging coup of his career: turning the troubled Howard Hughes Medical Institute into the richest research charity in the world.

Founded by the reclusive billionaire and aviator in 1953 as a tax shelter, the institute survived both Howard Hughes and myriad legal challenges. But in 1987, just when the institute seemed to have put its affairs in order, it was rocked by another scandal. The wife of its then president, Donald S. Fredrickson, had incurred some $200,000 in decorating expenses that wound up on the Hughes books; Fredrickson resigned. Then, in 1990, George W. Thorn, chairman of the board of trustees, announced his retirement.

The institute needed a leader with a flawless reputation and an uncommonly steady hand. It turned to Shapiro, a member of its board since 1984. Once again, he was at the helm of a multibillion-dollar organization founded on science, a field that by his own admission he knows little about. "I ducked out of every science class I ever took!" he exclaims mirthfully.

It was Shapiro's corporate management acumen, not his technical skills, that was at the root of his most significant contributions to Hughes Medical. As a board member, he engineered the sale of Hughes Aircraft to General Motors in 1985 for $5.2 billion, thus creating the largest private philanthropy in the U.S. with total assets of about $7 billion. (Shapiro abstained from the final selection because of his links with the bidders—he sat on the board of one, Boeing, and the other two, General Motors and Ford, were major customers of Du Pont.) But the complex sale was completed at almost exactly the market's peak. "We had very good fortune," Shapiro says with a professionally modest smile. Two years later Shapiro was instrumental in finally resolving a long-running battle with the Internal Revenue Service.

JOHN McGRAIL

***IRVING S. SHAPIRO** has run two major science-based concerns, although he "ducked out of every science class" he took.*

Now, after nearly three years of Shapiro's direct leadership, the Hughes Medical Institute may at last be finding its feet. "He leads the board with skill and dedication and in fine style," says trustee Alexander G. Bearn, an adjunct professor at the Rockefeller University and a former vice president of Merck Sharp & Dohme. "It's almost like a Quaker meeting. We have quite rigorous discussions, but we come to a consensus. I'd say it's a very happy board."

The most visible sign of Shapiro's second career was the dedication in May of a new $55-million administrative headquarters and conference center for Hughes scientists in Chevy Chase, Md. The institute also spent $281 million on medical research last year, an eightfold increase over the past decade. In the past few years it has inaugurated a series of initiatives for supporting science education. Its fellowship programs now extend to undergraduates as well as graduates, and it operates at liberal arts and traditionally black colleges as well as at schools that are primarily scientific. In addition, the institute has awarded grants to museums, health policy studies and several biomedical research groups. Expenditures on science education and other grants last year totaled $51.5 million; in all, the institute expects to spend $318 million this year.

For a biomedical investigator to receive an appointment at one of the 53 Hughes laboratories around the country is generally considered the next best thing to a meal ticket for life. The laboratories, which are associated with major research hospitals, are the institute's main focus, and the 223 Hughes investigators working at them include many of the top names in biology.

The Hughes approach—giving funds to exceptional scientists rather than to particular investigations—means that most of the work is fundamental in nature. Genetics and immunology, as well as cell biology and structural biology, are the favored areas. "We get many letters from people interested in a particular disease, asking, 'Can you do something? It will only take $3 million.' I understand their motivations, but that's not our way of looking at the world," Shapiro states firmly.

To judge from the number of important advances, the formula works. Since 1990 Hughes investigators have found the gene that is defective in neurofibromatosis, elucidated the structure of the protein that the AIDS virus uses to enter cells and identified genes associated with Lou Gehrig's disease and Huntington's disease, to note just a few examples. Recently Hughes has extended its support to researchers who choose to stay at their own institutions, broadening its impact still further.

The business expertise Shapiro now applies to biomedical research was first exercised when he was helping out in his family's clothes-pressing business in Minnesota during the 1930s. For his Lithuanian-born Jewish parents, "it was tough slogging in terms of money," he recalls. Shapiro escaped to the University of Minnesota, where he graduated in 1941 with a bachelor of law degree. He spent the next 10 years in Washington, D.C., in the criminal division of the Justice Department. Shapiro cites his experience in government as a source of his sense of corporate responsibility. "Franklin D. Roosevelt was God to the poor people in those days," he explains.

When Du Pont hired him in 1951, Shapiro says he felt obliged to point out his lack of scientific background. The company's general counsel replied, as Shapiro remembers it, that he was wanted not for his scientific knowledge but for his expertise in the law of conspiracy. During his time at the Justice Department, Shapiro had made an impression as a whiz in antitrust law. His experience served Du Pont well. Shapiro distinguished himself by being willing to take calculated risks, but he believes he was also recognized as a fair player. He climbed the ladder, becoming a vice president in 1970 and chairman and chief executive in 1974.

The appointment caused a sensation, in part because he was the first lawyer in the position but more because there were then few Jews in top-ranking jobs in corporate America. "Kingman Brewster [a former president of Yale University and ambassador to the Court of St. James] told me he would not have been surprised to hear there was a Jewish president of the U.S., but he was amazed to hear of a Jewish head of a major U.S. corporation," Shapiro recounts amiably. But Du Pont gave him a warm reception, and he notes with satisfaction that many prominent Jews have since told him he opened their door into the executive suite.

Shapiro understood from the start the importance of putting resources into research. To compensate for his personal unfamiliarity with scientific matters, he designated a member of Du Pont's executive committee to keep him informed, and he soon earned a reputation for decisive action. At the first shareholders' annual general meeting after he took charge as chairman, he announced that the company would be borrowing to finance a $2.5-billion investment in research and development. "It takes a fair amount of intellectual discipline to recognize that R&D is essential to prepare a company for one's successors," he reflects. "Any CEO who is not thinking long-term is tempted to cut."

Cooperation between business and the government "has slipped a little bit."

During the same period, he became an outspoken champion of corporate social responsibility as a founder of the Business Roundtable in 1972. The roundtable still exists as an association of business executives that, according to its literature, develops "positions which seek to reflect sound economic and social principles." Shapiro tried, he says, to "create the feeling that you have to be a constructive citizen whatever your private political beliefs."

He believes the efforts have had an enduring influence, especially easing the almost open warfare in the 1970s between business and an "absolutist" Environmental Protection Agency. But Shapiro sees troubling signs of deterioration in that entente cordiale. The Business Roundtable, he notes, is less closely identified with government-business cooperation than it was: "It has slipped a little bit," he says.

Shapiro's message secured him a position on a government advisory committee during the Carter years. There was a price to pay—during the Reagan terms, he says, he was twice approached about working for the executive branch, once to consult on defense reforms and once as an adviser on Middle East politics. But his ties with the Carter administration apparently proved too much for the White House. Even though years earlier he had taken his friend George Schultz on a tour of Israel and Jordan (which Shapiro says gave the future secretary of state "his education" about the region), the job offers mysteriously evaporated. "They blackballed me," he states matter-of-factly.

His talent was nonetheless sought by the legal profession. After retiring from Du Pont, he promptly took a partnership in the Wilmington office of Skadden, Arps, Slate, Meagher and Flom, a powerhouse law firm with offices in 12 countries. Then, in 1984, he became one of eight "prominent citizens" appointed by a Delaware judge as trustees to restore the Howard Hughes Medical Institute's affairs to order. Shapiro threw himself into rebuilding relations with officialdom. "The first thing we tried to do was end the running sore with the IRS—that was enough to turn your hair white," he grimaces.

Shapiro still retains links with Du Pont and displays his enthusiasm for industrial research on a wall of his law office overlooking the Delaware River. Behind his desk hangs a copy of a painting of the three Du Pont cousins, great-grandchildren of the company's founder, Eleuthère Irénée du Pont de Nemours, who bought the company in 1902 and turned it into a world leader. These days Shapiro says he is encouraged by the Clinton administration's stated intention of implementing a vigorous technology policy. "Industry has sometimes not taken research seriously enough," he muses. "The government might make mistakes, but I know we also made mistakes when I was at Du Pont."

Shapiro now spends four days a week at Skadden, Arps and one day a week on Hughes business. But he makes a point of going to the institute's scientific briefings, even though he does not follow many of the reports. "I go to put names and faces together," he explains. "There's a great value in letting scientists know who I am." When a researcher wrote him recently to take exception to a Hughes policy on intellectual property, "I called her up and said, 'Let's get together and talk,'" he says. "You can do a lot, assuming good faith."

Indeed, the institute has initiated a grants program to fund investigators in countries such as Mexico, Canada, New Zealand, Australia and Britain. Shapiro also sees a great opportunity beckoning in Eastern Europe and the former Soviet Union, where "$10,000 will buy you a lot of science." Plans for an initiative in those regions are well advanced.

Impatiently acknowledging a few of society's more crushing problems, Shapiro nonetheless predicts that the U.S. in the next century will be "history's richest society in quality of life." He finds his personal reward when he reads a popular account of some biomedical discovery and realizes "one of our people has really moved the ball forward, and it's because we supplied the money and picked the right person." *—Tim Beardsley*

SCIENTIFIC AMERICAN

Risk Analysis and Management

Inadequate approaches to handling risks may result in bad policy. Fortunately, rational techniques for assessment now exist

by M. Granger Morgan

Americans live longer and healthier lives today than at any time in their history. Yet they seem preoccupied with risks to health, safety and the environment. Many advocates, such as industry representatives promoting unpopular technology or Environmental Protection Agency staffers defending its regulatory agenda, argue that the public has a bad sense of perspective. Americans, they say, demand that enormous efforts be directed at small but scary-sounding risks while virtually ignoring larger, more commonplace ones.

Other evidence, however, suggests that citizens are eminently sensible about risks they face. Recent decades have witnessed precipitous drops in the rate and social acceptability of smoking, widespread shifts toward low-fat, high-fiber diets, dramatic improvements in automobile safety and the passage of mandatory seat belt laws—all steps that reduce the chance of untimely demise at little cost.

My experience and that of my colleagues indicate that the public can be very sensible about risk when companies, regulators and other institutions give it the opportunity. Laypeople have different, broader definitions of risk, which in important respects can be more rational than the narrow ones used by experts. Furthermore, risk management is, fundamentally, a question of values. In a democratic society, there is no acceptable way to make these choices without involving the citizens who will be affected by them.

The public agenda is already crowded with unresolved issues of certain or potential hazards such as AIDS, asbestos in schools and contaminants in food and drinking water. Meanwhile scientific and social developments are bringing new problems—global warming, genetic engineering and others—to the fore. To meet the challenge that these issues pose, risk analysts and managers will have to change their agenda for

M. GRANGER MORGAN has worked for many years to improve techniques for analyzing and managing risks to health, safety and the environment. Morgan heads the department of engineering and public policy at Carnegie Mellon University. He also holds appointments in the department of electrical and computer engineering and at the H. John Heinz III School of Public Policy and Management. Morgan received a B.A. from Harvard University, an M.S. from Cornell University and a Ph.D. in applied physics from the University of California, San Diego.

AIR DISASTER in Madrid claimed 183 lives in November 1983. The (small) chance of dying in an air crash is one of the prices that society agrees to pay for rapid, convenient global transportation. Some risks, including nuclear power generation, have caused fewer deaths but provoked greater calls for regulation, whereas others, such as automobiles, cause more deaths but arouse less concern.

evaluating dangers to the general welfare; they will also have to adopt new communication styles and learn from the populace rather than simply trying to force information on it.

While public trust in risk management has declined, ironically the discipline of risk analysis has matured. It is now possible to examine potential hazards in a rigorous, quantitative fashion and thus to give people and their representatives facts on which to base essential personal and political decisions.

Risk analysts start by dividing hazards into two parts: exposure and effect. Exposure studies look at the ways in which a person (or, say, an ecosystem or a piece of art) might be subjected to change; effects studies examine what may happen once that exposure has manifested itself. Investigating the risks of lead for inner-city children, for example, might start with exposure studies to learn how old, flaking house paint releases lead into the environment and how children build up the substance in their bodies by inhaling dust or ingesting dirt. Effects studies might then attempt to determine the reduction in academic performance attributable to specific amounts of lead in the blood.

Exposure to a pollutant or other hazard may cause a complex chain of events leading to one of a number of effects, but analysts have found that the overall result can be modeled by a function that assigns a single number to any given exposure level. A simple, linear relation, for instance, accurately describes the average cancer risk incurred by smokers: 10 cigarettes a day generally increase the chance of contracting lung cancer by a factor of 25; 20 cigarettes a day increase it by a factor of 50. For other risks, however, a simple dose-response function is not appropriate, and more complex models must be used.

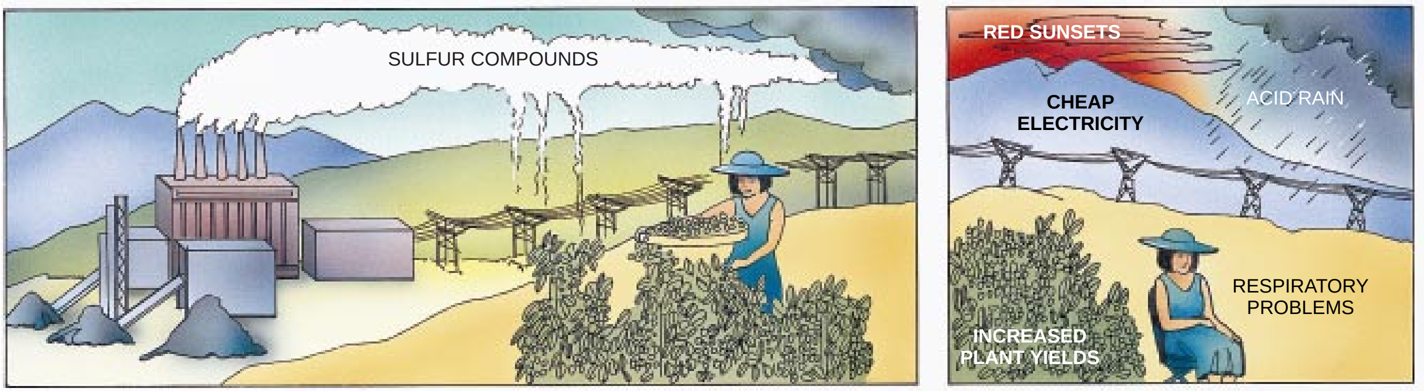
The study of exposure and effects is fraught with uncertainty. Indeed, uncertainty is at the heart of the definition of risk. In many cases, the risk may be well understood in a statistical sense but still be uncertain at the level of individual events. Insurance companies cannot predict whether any single driver will be killed or injured in an accident, even though they can estimate the annual number of crash-related deaths and injuries in the U.S. with considerable precision.

For other risks, such as those involving new technologies or those in which bad outcomes occur only rarely, uncertainty enters the calculations at a higher level—overall probabilities as well as individual events are unpredictable. If good actuarial data are not available, analysts must find other methods to estimate the likelihood of exposure and subsequent effects. The development of risk assessment during the past two decades has been in large part the story of finding ways to determine the extent of risks that have little precedent.

In one common technique, failure mode and effect analysis, workers try to identify all the events that might help cause a system to break down. Then they compile as complete a description

EXPOSURE PROCESSES

EFFECTS PROCESSES



RISK MANAGEMENT PROCESS begins with analysis of the people and other entities exposed to change, such as in this illustration, from emissions from a coal-burning power plant (*left*). After the results of exposure have been quantified (*second panel*), they must then be filtered through public perceptions, which cause people to respond more strongly to some

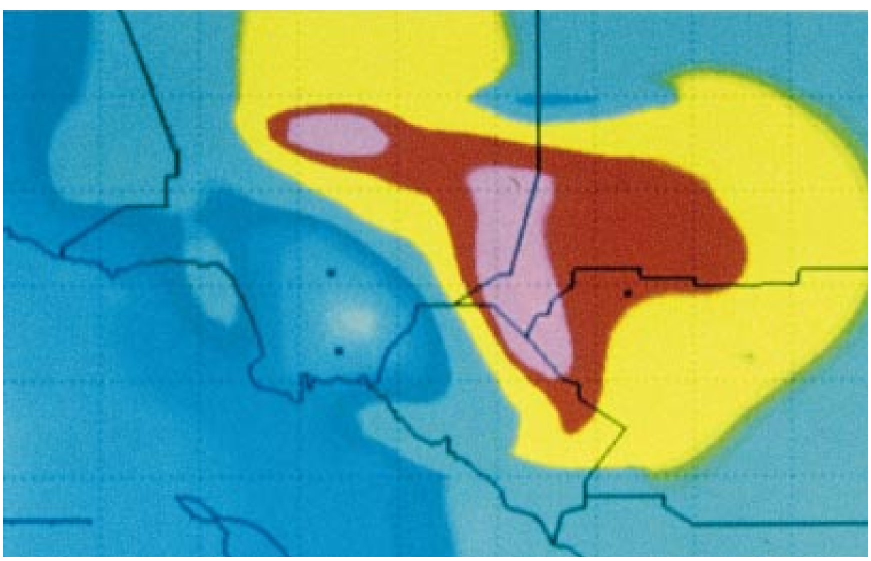
as possible of the routes by which those events could lead to a failure (for instance, a chemical tank might release its contents either because a weld cracks and the tank ruptures or because an electrical short causes the cooling system to stop, allowing the contents to overheat and eventually explode). Although enumerating all possible routes to failure may sound like a simple task, it is difficult to exhaust all the alternatives. Usually a system must be described several times in different ways before analysts are confident that they have grasped its intricacies, and even then it is often impossible to be sure that all avenues have been identified.

Once the failure modes have been enumerated, a fault tree can aid in estimating the likelihood of any given mode. This tree graphically depicts how the subsystems of an object depend on one another and how the failure of one part affects key operations. Once the fault tree has been constructed, one need only estimate the probability that individual elements will fail to find the chance that the entire system will cease to function under a particular set of circumstances. Norman C. Rasmussen of the Massachusetts Institute of Technology was among the first to use the method on a large scale when he directed a study of nuclear reactor safety in 1975. Although specific details of his estimates were disputed, fault trees are now used routinely in the nuclear industry and other fields.

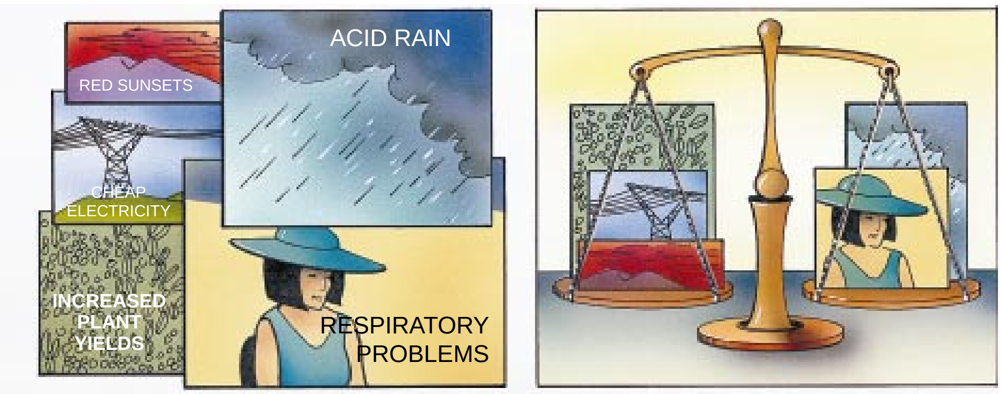
Boeing applies fault-tree analysis to the design of large aircraft. Company engineers have identified and remedied a number of potential problems, such as vulnerabilities caused by routing multiple control lines through the same area. Alcoa workers recently used fault trees to examine the safety of their large furnaces. On the basis of their findings, the company revised its safety standards to mandate the use of programmable logic controllers for safety-critical controls. They also instituted rigorous testing of automatic shut-off valves for leaks and added alarms that warn operators to close manual isolation valves during shutdown periods. The company estimates that these changes have reduced the likelihood of explosions by a factor of 20. Major chemical companies such as Du Pont, Monsanto and Union Carbide have also employed the technique in designing processes for chemical plants, in deciding where to build plants and in evaluating the risks of transporting chemicals.

In addition to dealing with uncertainty about the likelihood of an event such as the breakdown of a crucial piece of equipment, risk analysts must cope with other unknowns: if a chemical tank leaks, one cannot determine beforehand the exact amount of pollutant released, the precise shape of the resulting dose-response curves for people exposed, or the values of the rate constants governing the chemical reactions that convert the contents of the tank to more or less dangerous forms. Such uncertainties are often represented by means of probability distributions, which describe the odds that a quantity will take on a specific value within a range of possible levels.

When risk specialists must estimate the likelihood that a part will fail or assign a range of uncertainty to an essential value in a model, they can sometimes use data collected from similar systems elsewhere—although the design of a proposed chemical plant as a whole may be new, the components in its high-pressure steam systems will ba-



SUPERCOMPUTER MODEL of ozone concentrations in the Los Angeles basin (*pink, highest; yellow, lowest*) serves as a starting point for analyses of the risks of exposure to air pollutants.



aspects of risk than to others. Ultimately, costs and benefits will be weighed. Agreeing on the values used to make decisions and making sure that all relevant effects are taken into account are crucial, but often neglected, parts of the process.

sically be indistinguishable from those in other plants.

In other cases, however, historical data are not available. Sometimes workers can build predictive models to estimate probabilities based on what is known about roughly similar systems, but often they must rely on expert subjective judgment. Because of the way people think about uncertainty, this approach may involve serious biases. Even so, quantitative risk analysis retains the advantage that judgments can be incorporated in a way that makes assumptions and biases explicit.

Only a few years ago such detailed study of risks required months of custom programming and days or weeks of mainframe computer time. Today a variety of powerful, general-purpose tools are available to make calculations involving uncertainty. These programs, many of which run on personal computers, are revolutionizing the field. They enable accomplished analysts to complete projects that just a decade ago were considered beyond the reach of all but the most sophisticated organizations [*see box on page 38*]. Although using such software requires training, they could democratize risk assessment and make rigorous determinations far more widely available.

After they have determined the likelihood that a system could expose people to harm and described the particulars of the damage that could result from exposure, some risk analysts believe their job is almost done. In fact, they have just completed the preliminaries. Once a risk has been identified and analyzed, psychological and social processes of perception and valuation come into play. How people view and evaluate particular risks determines which of the many changes that may occur in the world they choose to notice and perhaps do something about. Someone must then establish the rules for weighing risks, for deciding if the risk is to be controlled and, if so, how. Risk management thus tends to force a society to consider what it cares about and who should bear the burden of living with or mitigating a problem once it has been identified.

For many years, most economists and technologists perceived risk simply in terms of expected value. Working for a few hours in a coal mine, eating peanut butter sandwiches every day for a month, and living next to a nuclear power plant for five years all involve an increased risk of death of about one in a million, so analysts viewed them all as equally risky. When people are asked to rank various activities and technologies in terms of risk, however, they produce lists whose order does not correspond very closely to the number of expected deaths. As a result, some early risk analysts decided that people were confused and that their opinions should be discounted.

Since then, social scientists have conducted extensive studies of public risk perception and discovered that the situation is considerably more subtle. When people are asked to order well-known hazards in terms of the number of deaths and injuries they cause every year, on average they can do it pretty well. If, however, they are asked to rank those hazards in terms of risk, they produce quite a different order.

People do not define risk solely as the expected number of deaths or injuries per unit time. Experimental psychologists Baruch Fischhoff of Carnegie Mellon University and Paul Slovic and Sarah Lichtenstein of Decision Research in Eugene, Ore., have shown that people also rank risks based on how well the process in question is understood, how equitably the danger is distributed, how well individuals can control their exposure and whether risk is assumed voluntarily.

Slovic and his colleagues have found that these factors can be combined into three major groups. The first is basically an event's degree of dreadfulness (as determined by such features as the scale of its effects and the degree to which it affects "innocent" bystanders). The second is a measure of how well the risk is understood, and the third is the number of people exposed. These groups of characteristics can be used to define a "risk space." Where a hazard falls within this space says quite a lot about how people are likely to respond to it. Risks carrying a high level of "dread," for example, provoke more calls for government intervention than do some more workaday risks that actually cause more deaths or injuries.

In making judgments about uncertainty, including ones about risk, experimental psychologists have found that people unconsciously use a number of heuristics. Usually these rules of thumb work well, but under some circumstances they can lead to systematic bias or other errors. As a result, people tend to underestimate the frequency of very common causes of death—stroke, cancer, accidents—by roughly a factor of 10. They also overestimate the frequency of very uncommon causes of death (botulism poisoning, for example) by as much as several orders of magnitude.

These mistakes apparently result from the so-called heuristic of availability. Daniel Kahneman of the University of California at Berkeley, Amos N. Tversky of Stanford University and others have found that people often judge the likelihood of an event in terms of how easily they can recall (or imagine) examples. In this case, stroke is a very common cause of death, but most people learn about it only when a close friend or relative or famous person dies; in contrast, virtually every time someone dies of botulism, people are likely to hear about it on the evening news. This heuristic and others are not limited to the general public. Even experts sometimes employ them in making judgments about uncertainty.

Once people have noticed a risk and decided that they care enough to do something about it, just what should they do? How should they decide the amount to be spent on reducing the risk, and on whom should they place the primary burdens? Risk managers can intervene at many points: they can work to prevent the process producing the risk, to reduce exposures, to modify effects, to alter perceptions or valuations through education and public relations or to compensate for damage after the fact. Which strategy is best depends in large part on the attributes of the particular risk.

Even before determining how to intervene, risk managers must choose the rules that will be used to judge whether to deal with a particular issue and, if so, how much attention, effort and money to devote. Most rules fall into one of three broad classes: utility based, rights based and technology based. The first kind of rules attempt to maximize net benefits. Analysts add up the pros and cons of a particular course of action and take the difference between the two. The course with the best score wins.

Early benefit-cost analyses employed fixed estimates of the value of good and bad outcomes. Many workers now use probabilistic estimates instead to reflect the inherent uncertainty of their descriptions. Although decisions are ultimately made in terms of expected values, other measures may be employed as well. For example, if the principal concern is to avoid disasters, analysts could adopt a "minimax" criterion, which seeks to minimize the harm done by the worst possible outcome, even if that leads to worse results on average.

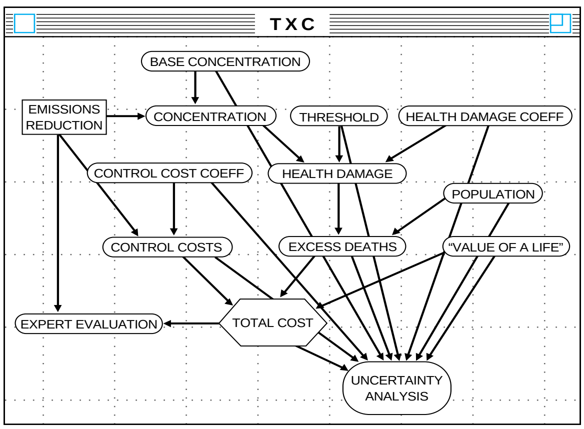
Of course, many tricky points are involved in such calculations. Costs and benefits may not depend linearly on the amount of pollutant emitted or on the number of dollars spent for control. Furthermore, not all the pros and cons of an issue can necessarily be measured on the same scale. When the absolute magnitude of net benefits cannot be estimated, however, rules based on relative criteria such as cost-effectiveness can still aid decision makers.

Rights-based rules replace the notion of utility with one of justice. In most utility-based systems, anything can be subject to trade-offs; in rights-based ones, however, there are certain things that one party cannot do to another without its consent, regardless of costs or benefits. This is the approach that Congress has taken (at least formally) in the Clean Air Act of 1970: the law does not call for maximizing net social benefit; instead it just requires controlling pollutant concentrations so as to protect the most sensitive populations exposed to them. The underlying pre-

Risk Analysis in Action

Uncertainty is a central element of most problems involving risk. Analysts today have a number of software tools that incorporate the effects of uncertainty. These tools can show the logical consequences of a particular set of risk assumptions and rules for making decisions about it. One such system is Demos, developed by Max Henrion of Lumina Decision Systems in Palo Alto, Calif.

To see how the process works, consider a hypothetical chemical pollutant, "TXC." To simplify matters, assume that the entire population at risk (30 million people) is exposed to the same dose—this makes a model of exposure processes unnecessary. The next step is to construct a function that describes the risk associated with any given exposure level—for example, a linear dose-response function, possibly with a threshold below which there is no danger.



BLOCKS in the diagram above can be expanded to call up a window containing graphs and tables for their assumptions, equations and probability distributions.

Given this information, Demos can estimate the number of excess deaths caused every year by TXC exposure. According to the resulting cumulative probability distribution, there is about a 30 percent chance that no one dies, about a 50 percent chance that fewer than 100 people die each year and about a 10 percent chance that more than 1,000 die.

Meanwhile, for a price, pollution controls can reduce the concentration of TXC. (The cost of achieving any given reduction, like the danger of exposure, is determined by consultation with experts.) To choose a level of pollution control that minimizes total social costs, one must first decide how much society is willing to invest to prevent mortality. The upper and lower bounds in this example are $300,000 and $3 million per death averted. (Picking such numbers is a value judgment; in practice, a crucial part of the analysis would be to find out how sensitive the results are to the dollar values placed on life or health.)

Net social costs, in this model, are simply the sum of control costs and mortality. At $300,000 per death averted, their most likely value reaches a minimum when TXC emissions are reduced by 55 percent. At $3 million, the optimum reduction is about 88 percent.

Demos can also calculate a form of correlation between each of the input variables and total costs. Strong correlations indicate variables that contribute significantly to the uncertainty in the final cost estimate. At low levels of pollution control, possible variations in the slope of the damage function, in the location of the threshold and in the base concentration of the pollutant contribute the most to total uncertainty. At very high levels of control, in contrast, almost all the uncertainty derives from unknowns in the cost of controlling emissions.

Finally, Demos can compute the difference in expected cost between the optimal decision based on current information and that given perfect information—that is, the benefit of removing all uncertainties from the calculations. This is known in decision analysis as the expected value of perfect information; it is an upper bound on the value of research. If averting a single death is worth $300,000 to society, this value is $38 million a year; if averting a death is worth $3 million, it is $71 million a year.

Although tools such as Demos put quantitative risk analysis within reach of any group with a personal computer, using them properly requires substantial education. My colleagues and I found that a group of first-year engineering doctoral students first exposed to Demos tended to ignore possible correlations among variables, thus seriously overestimating the uncertainty of their results.

sumption holds that these individuals have a right to protection from harm.

Technology-based criteria, in contrast to the first two types, are not concerned with costs, benefits or rights but rather with the level of technology available to control certain risks. Regulations based on these criteria typically mandate "the best available technology" or emissions that are "as low as reasonably achievable." Such rules can be difficult to apply because people seldom agree on the definitions of "available" or "reasonably achievable." Furthermore, technological advances may impose an unintended moving target on both regulators and industry.

There is no correct choice among the various criteria for making decisions about risks. They depend on the ethical and value preferences of individuals and society at large. It is, however, critically important that decision frameworks be carefully and explicitly chosen and that these choices be kept logically consistent, especially in complex situations. To do otherwise may produce inconsistent approaches to the same risk. The EPA has slipped into this error by writing different rules to govern exposure to sources of radioactivity that pose essentially similar risks.

Implicit in the process of risk analysis and management is the crucial role of communication. If public bodies are to make good decisions about regulating potential hazards, citizens must be well informed. The alternative of entrusting policy to panels of experts working behind closed doors has proved a failure, both because the resulting policy may ignore important social considerations and because it may prove impossible to implement in the face of grass-roots resistance.

Until the mid-1980s, there was little research on communicating risks to the public. Over the past five years, along with my colleagues Fischhoff and Lester B. Lave, I have found that much of the conventional wisdom in this area does not hold up. The chemical industry, for example, distilled years of literature about communication into advice for plant managers on ways to make public comparisons between different kinds of risks. We subjected the advice to empirical evaluation and found that it is wrong. We have concluded that the only way to communicate risks reliably is to start by learning what people already know and what they need to know, then develop messages, test them and refine them until surveys demonstrate that the messages have conveyed the intended information.

In 1989 we looked at the effects of the EPA's general brochure about radon in homes. The EPA prepared this brochure according to traditional methods: ask scientific experts what they think people should be told and then package the result in an attractive form. In fact, people are rarely completely ignorant about a risk, and so they filter any message through their existing knowledge. A message that does not take this filtering process into account can be ignored or misinterpreted.

To study people's mental models, we began with a set of open-ended interviews, first asking, "Tell me about radon." Our questions grew more specific only in the later stages of the interview. The number of new ideas encountered in such interviews approached an asymptotic limit after a couple of dozen people. At this point, we devised a closed-form questionnaire from the results of the interviews and administered it to a much larger sample.

We uncovered critical misunderstandings in beliefs that could undermine the effectiveness of the EPA's messages. For example, a sizable proportion of the public believes that radon contamination is permanent and does not go away. This misconception presumably results from an inappropriate inference based on knowledge about chemical contaminants or long-lived radioisotopes. The first version of the EPA's "Citizen's Guide to Radon" did not discuss this issue. Based in part on our findings, the latest version addresses it explicitly.

The objective of risk communication is to provide people with a basis for making an informed decision; any effective message must contain information that helps them in that task. With former doctoral students Ann Bostrom, now at the Georgia Institute of Technology, and Cynthia J. Atman, now at the University of Pittsburgh, we used our method to develop two brochures about radon and compared their effectiveness with that of the EPA's first version. When we asked people to recall simple facts, they did equally well with all three brochures. But when faced with tasks that required inference—advising a neighbor with a high radon reading on what to do—people who received our literature dramatically outperformed those who received the EPA material.

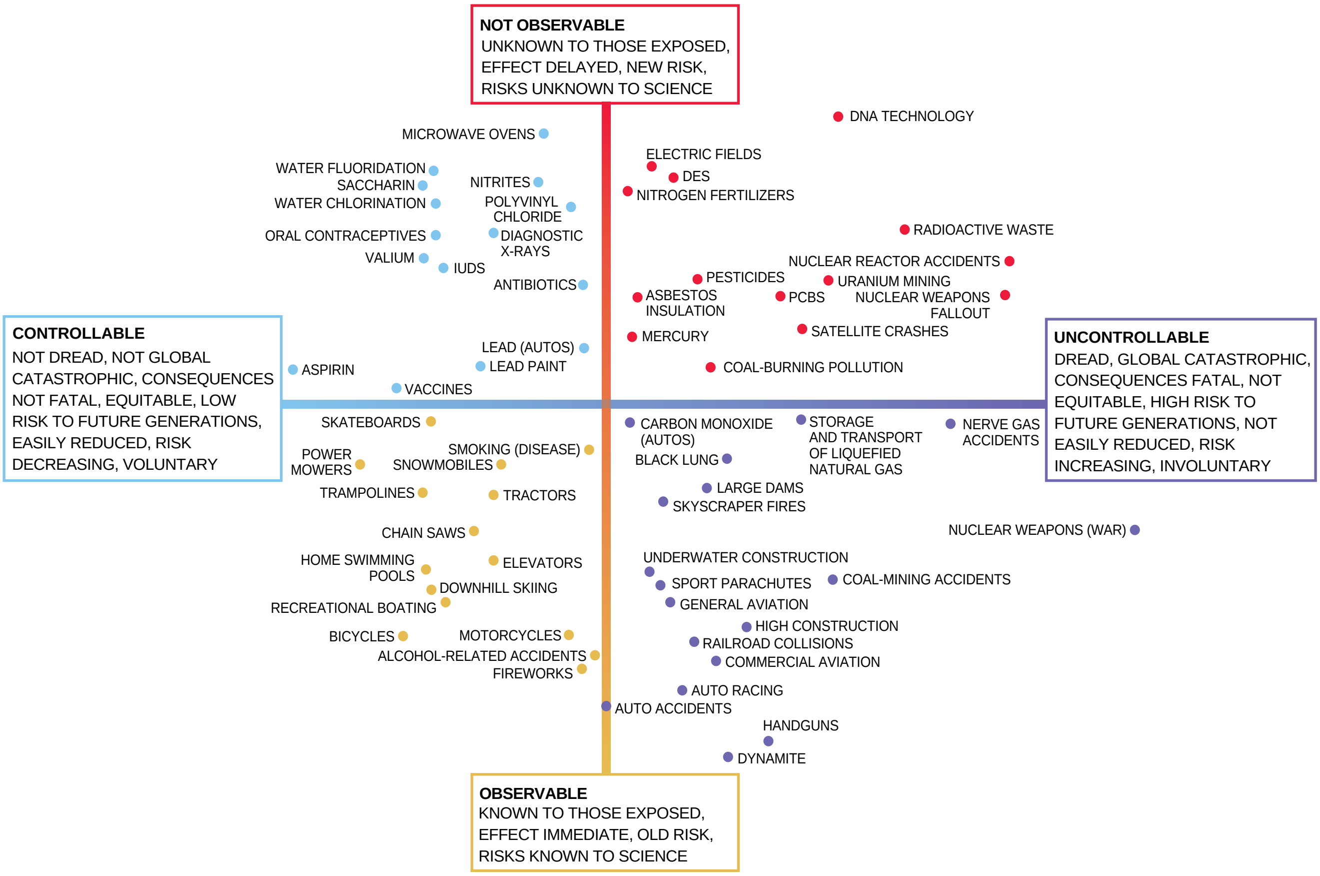
We have found similar misperceptions in other areas, say, climatic change. Only a relatively small proportion of people associate energy use and carbon dioxide emissions with global warming. Many believe the hole in the ozone layer is the factor most likely to lead to global warming, although in fact the two issues are only loosely connected. Some also think launches of spacecraft are the major contributor to holes in the ozone layer. (Willett Kempton of the University of Delaware has found very similar perceptions.)

The essence of good risk communication is very simple: learn what people already believe, tailor the communication to this knowledge and to the decisions people face and then subject the resulting message to careful empirical evaluation. Yet almost no one communicates risks to the public in this fashion. People get their information in fragmentary bits through a press that often does not understand technical details and often chooses to emphasize the sensational. Those trying to convey information are generally either advocates promoting a particular agenda or regulators who sometimes fail either to do their homework or to take a sufficiently broad perspective on the risks they manage. The surprise is not that opinion on hazards may undergo wide swings or may sometimes force silly or inefficient outcomes. It is that the public does as well as it does.

Indeed, when people are given balanced information and enough time to reflect on it, they can do a remarkably good job of deciding what problems are important and of systematically addressing decisions about risks. I conducted studies with Gordon Hester (then a doctoral student, now at the Electric Power Research Institute) in which we asked opinion leaders—a teacher, a state highway patrolman, a bank manager and so on—to play the role of a citizens' board advising the governor of Pennsylvania on the siting of high-voltage electric transmission lines. We asked the groups to focus particularly on the controversial problem of health risks from electric and magnetic fields emanating from transmission lines. We gave them detailed background information and a list of specific questions. Working mostly on their own, over a period of about a day and a half (with pay), the groups structured policy problems and prepared advice in a fashion that would be a credit to many consulting firms.

If anyone should be faulted for the poor quality of responses to risk, it is probably not the public but rather risk managers in government and industry. First, regulators have generally adopted a short-term perspective focused on taking action quickly rather than investing in the research needed to improve understanding of particular hazards in the future. This focus is especially evident in regulations that have been formulated to ensure the safety of the environment, workplace and consumer products.

Second, these officials have often



RISK SPACE has axes that correspond roughly to a hazard's "dreadfulness" and to the degree to which it is understood. Risks in the upper right quadrant of this space are most likely to provoke calls for government regulation.

adopted too narrow an outlook on the risks they manage. Sometimes attempts to reduce one risk (burns from flammable children's pajamas) have created others (the increased chance of cancer from fireproofing chemicals).

In some instances, regulators have ignored large risks while attacking smaller ones with vigor. Biologist Bruce Ames of Berkeley has argued persuasively that government risk managers have invested enormous resources in controlling selected artificial carcinogens while ignoring natural ones that may contribute far more to the total risk for human cancer.

Third, government risk managers do not generally set up institutions for learning from experience. Too often adversarial procedures mix attempts to figure out what has happened in an incident with the assignment of blame. As a result, valuable safety-related insights may either be missed or sealed away from the public eye. Civilian aviation, in contrast, has benefited extensively from accident investigations by the National Transportation Safety Board. The board does its work in isolation from arguments about liability; its results are widely published and have contributed measurably to improving air safety.

Many regulators are probably also too quick to look for single global solutions to risk problems. Experimenting with multiple solutions to see which ones work best is a strategy that deserves far more attention than it has received. With 50 states in a federal system, the U.S. has a natural opportunity to run such experiments.

Finally, risk managers have not been sufficiently inventive in developing arrangements that permit citizens to become involved in decision making in a significant and constructive way, working with experts and with adequate time and access to information. Although there are provisions for public hearings in the licensing process for nuclear reactors or the siting of hazardous waste repositories, the process rarely allows for reasoned discussion, and input usually comes too late to have any effect on the set of alternatives under consideration.

Thomas Jefferson was right: the best strategy for assuring the general welfare in a democracy is a well-informed electorate. If the U.S. and other nations want better, more reasoned social decisions about risk, they need to take steps to enhance public understanding. They must also provide institutions whereby citizens and their representatives can devote attention to risk management decisions. This will not preclude the occasional absurd outcome, but neither does any other way of making decisions. Moreover, appropriate public involvement should go a long way toward eliminating the confrontational tone that has become so common in the risk management process.

FURTHER READING

RATIONAL CHOICE IN AN UNCERTAIN WORLD. Robyn M. Dawes. Harcourt Brace Jovanovich, 1988.

READINGS IN RISK. Edited by Theodore S. Glickman and Michael Gough. Resources for the Future, 1990.

UNCERTAINTY: A GUIDE TO DEALING WITH UNCERTAINTY IN QUANTITATIVE RISK AND POLICY ANALYSIS. M. Granger Morgan and Max Henrion. Cambridge University Press, 1990.

COMMUNICATING RISK TO THE PUBLIC. M. Granger Morgan, Baruch Fischhoff, Ann Bostrom, Lester Lave and Cynthia J. Atman in *Environmental Science and Technology,* Vol. 26, No. 11, pages 2048–2056; November 1992.

RISK ANALYSIS. Publication of the Society for Risk Analysis, published quarterly by Plenum Publishing.

Viral Quasispecies

The standard definition of a biological species does not apply to viruses. A more expansive and dynamic view of viral populations holds clues to understanding and defeating them

by Manfred Eigen

According to Greek mythology, when curious Pandora opened a forbidden box she set loose all the miseries and evils known to the world. One of them was undoubtedly the virus—the very name of which is Latin for slime, poison and stench. Viruses cause a mind-boggling assortment of illnesses, ranging from the common cold to acquired immunodeficiency syndrome (AIDS), perhaps the most feared scourge of modern times.

Viruses have the ability to mystify laypeople and experts alike. Early in their studies of viruses, investigators became puzzled by the high mutation rates they observed: the magnitudes indicated that viruses must evolve more than a million times faster than cellular microorganisms. If that were true, how could viruses maintain their identities as pathogenic species over any evolutionarily significant period? Why didn't they mutate out of existence?

Those questions have generally been unanswerable within the traditional theoretical framework of biology. Borrowing ideas from both mathematics and chemistry, however, my colleagues and I have recently introduced a concept, the quasispecies, that can illuminate the problems in new ways. A viral species, we have shown, is actually a complex, self-perpetuating population of diverse, related entities that act as a whole.

The substitution of "quasispecies" for "species" is not merely semantic. It offers insights into the behavior of viruses. In the case of AIDS, for example, it helps in determining when the human immunodeficiency virus (HIV) first evolved and where it may have come from. If one were to extrapolate only from the epidemiologic data, AIDS would seem to have first appeared in 1979. Our data, in contrast, suggest that HIV is a very old virus. Moreover, the quasispecies concept points toward potential treatments for AIDS and other diseases that have so far been resistant to vaccines.

MANFRED EIGEN is director of biochemical kinetics research at the Max Planck Institute for Biophysical Chemistry in Göttingen, where he began his undergraduate studies in 1951. For his ground-breaking work in developing techniques for measuring high-speed chemical reactions, Eigen was named as a co-recipient of the 1967 Nobel Prize for Chemistry. In more recent years the major focus of his research has been the significance of the information concept to molecular evolution and its technological applications.

To begin to understand viral quasispecies, we must ask ourselves, What is a virus? In 1959 Nobel laureate André Lwoff's answer was "A virus is a virus!"—a truism, perhaps, but one that cuts to the uniqueness of viruses in the living world. Essentially, a virus is a genetic program that carries the simple message "Reproduce me!" from one cell to another. Because a virus represents only one or a few of the messengers vying for the attention of its host, it must employ certain biochemical tricks to recruit the host's replication machinery for its selfish purpose. Often those ploys result in the host cell's death.

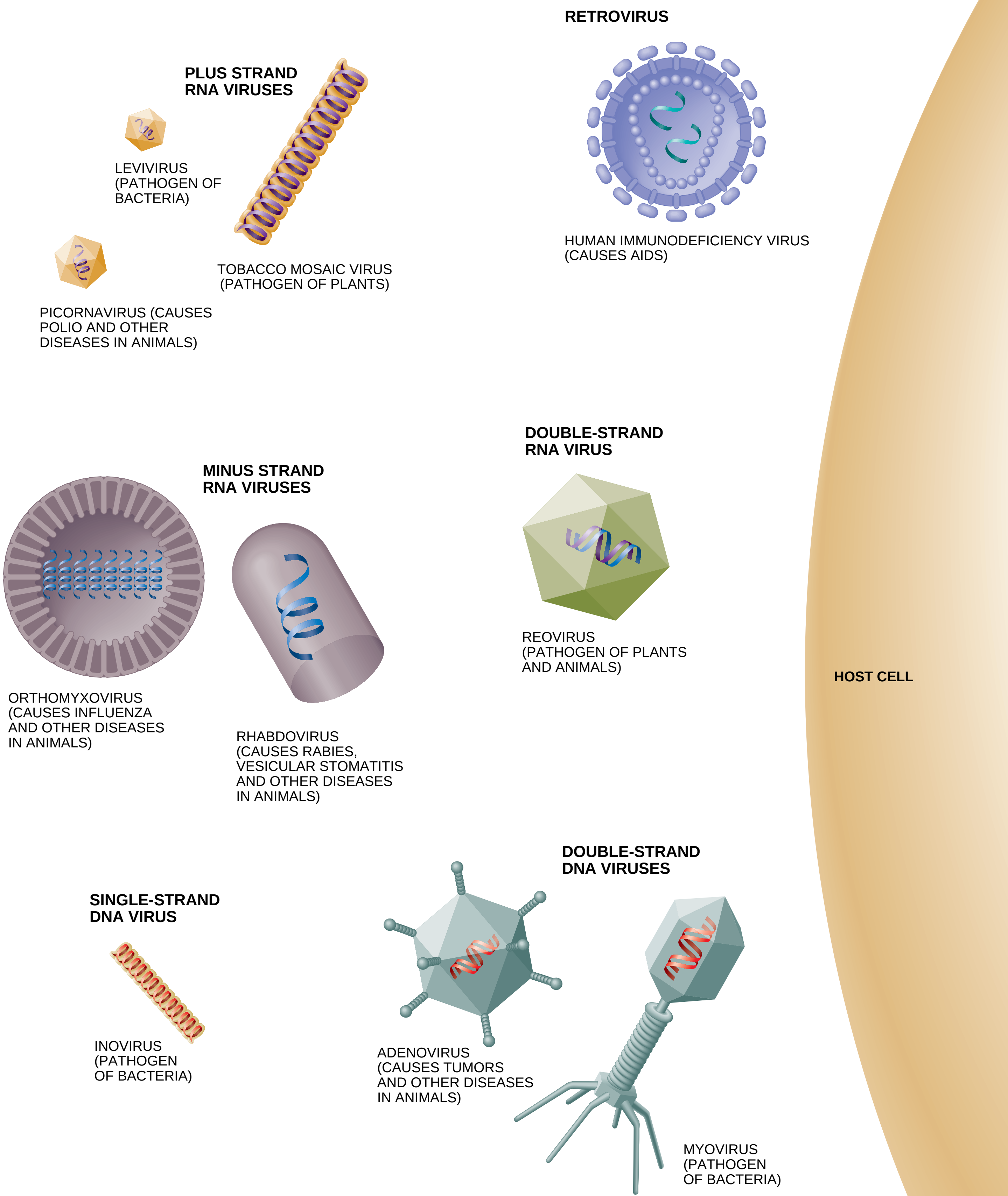
Viruses fall into many different categories, but one way to distinguish among them is by looking at the molecules that carry their genetic messages. Perhaps the simplest form of virus is represented by a single strand of ribonucleic acid (RNA), made up of several thousand individual nucleotide subunits. If this RNA is a so-called plus strand, it can be read directly by the host's translation apparatus, the ribosome, much as the host's own messenger RNA can. Examples of such plus strand viruses are the bacteriophage Qß, a parasite of the bacterium *Escherichia coli,* and the polio-1 virus, which causes spinomuscular paralysis. Other viruses encode their messages as minus strands of RNA. Inside a cell, minus strands must be transcribed into complementary plus strands before viral replication can begin. Influenza A, one of the most common epidemic diseases, is caused by a minus strand virus.

A third class of single-strand RNA viruses consists of retroviruses. After a retrovirus infects a host cell, a viral enzyme called reverse transcriptase changes the single strand of viral RNA into a double strand of deoxyribonucleic acid (DNA). That DNA can then incorporate itself into the host's genome, thereby making the viral message an inheritable feature of the cell. HIV belongs to the retroviral family. Its target is the immune system, which ought to provide protection against the virus.

Because viruses are so dependent on the replicative systems of their hosts, scientists generally believe viruses in their present form must have evolved after cellular life. It is even possible that viruses descended from parts of their host's genetic programs that turned their inside knowledge of cells to the goal of duplicating themselves. Whatever the case, viruses are useful models for studying how molecules may have organized themselves into self-perpetuating units at the dawn of life. They show how information can be generated and processed at the molecular level. The essence of their genetic information is self-preservation, which they achieve through mutagenesis, reproduction, proliferation and adaptation to a steadily changing environment.

The genome of a single-strand RNA virus such as HIV, which comprises only 10,000 nucleotides, is small and simple compared with that of most cells. Yet from a molecular standpoint, it is unimaginably complex. Each of those nucleotides contains one of four possible bases: adenine, uracil, guanine or cytosine. The unique sequence specified by the genome of HIV therefore represents just one choice out of $4^{10,000}$ possibilities—a number roughly equivalent to a one followed by 6,000 zeros.

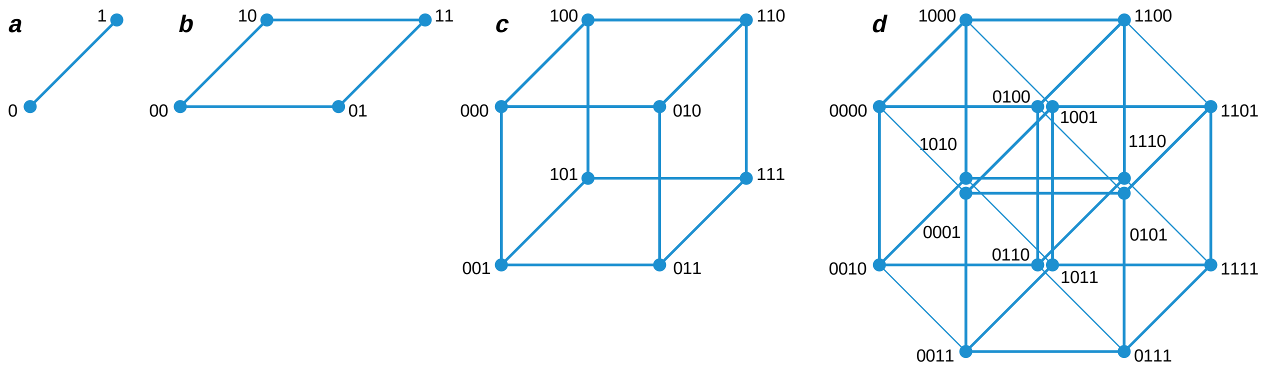
Most such sequences would not qualify as viruses: they could not direct



VIRUSES BELONG to many diverse families, which may be distinguished by the type and activities of their genetic molecules. In some viruses the genes are in single or double strands of DNA; in others the genes are RNA molecules. Some RNA viruses carry plus strands that can be translated directly by the host cell's protein-making machinery. For minus strand viruses, the RNA must first be transcribed into complementary plus strands. Retroviruses, such as those that cause AIDS, require that their RNA be reverse-transcribed into double strands of DNA. Only a few of the many varieties of viruses are shown.

How to Construct a Sequence Space

One way to study the diverse nucleotide sequences in the genes of viruses is to map them into a multidimensional matrix called a Hamming sequence space. In this space, each point represents a unique sequence, and the degree of separation between points reflects their degree of dissimilarity. The space can be most easily drawn for short sequences consisting of binary digits. For a sequence with just one position, there are only two possible sequences, and they can be drawn as the end points of a line (*a*). For a sequence with two positions, there are four permutations, which form the corners of a square (*b*). The variations on a three-digit sequence become the corners of a cube (*c*), and the variations on a four-digit sequence are the vertices of a four-dimensional hypercube (*d*). Each higher-dimensional space is built iteratively by drawing the previous diagram twice and connecting the corresponding points. The sequence spaces for viral genomes are far more complex than these simple figures because they involve thousands of positions that can each be occupied by one of four different nucleotides.



their own duplication. Nevertheless, even if only a tiny fraction of them are viruses, the number is still huge. If the entire universe were completely filled with hydrogen atoms—each about one trillionth of a trillionth of a cubic centimeter in volume—it could hold only about 10^{108} of them. Hence, an array of $10^{6,000}$ differing RNA sequences is beyond comprehension.

Fortunately, it is not beyond the analytic reach of mathematics. We can construct a theoretical framework that encompasses that vast array and reveals relations among the elements. To do so, we must first develop a geometry—a concept of space—that would allow us to represent the informational differences among the sequences as precise spatial distances. In this space, each nucleotide sequence must occupy a unique position. The positions must also be arranged to reflect the informational kinship between the sequences. In other words, each sequence should be only one unit away from all the other sequences that differ from it by only one nucleotide; it should be two units away from those differing by two nucleotides, and so on.

Sequence space proves to be an invaluable tool for interpreting what a viral species is. The term "species" is used in both biology and chemistry. In chemistry, a species is a defined chemical compound, such as trinitrotoluene or benzene. In biology, the definition is not quite as sharp: members of a given living species must show common traits and must be at least potentially able to produce offspring by recombining their genetic material. At the genetic level, a biological species is represented by a gigantic variety of differing DNA molecules.

Biologists generally speak of the wild type of a species: the form that predominates in a population and that is particularly well suited to the environment in which it lives. If one found an individual that perfectly embodied that wild type, its unique sequence of genomic DNA would specify the wild type at the genetic level and would occupy a single point in the sequence space. That view of the wild type accords with the classical model of natural selection. Although mutations occur steadily, they presumably disappear because the mutant types are less fit than the wild type. Alternatively, a mutant may have advantages, in which case it becomes the new wild type. Either outcome tends to keep all the members of a species at or very near one point in a genome sequence space.

That picture was modified by the neutral theory advanced in the 1960s by Motoo Kimura of the National Institute of Genetics in Mishima, Japan. Kimura argued that many mutations, such as those causing differences in blood types, are neither advantageous nor disadvantageous. Consequently, a small but statistically defined fraction of the neutral mutations would continuously replace the existing wild type in the population. The genome of a species would therefore drift steadily but randomly through a certain volume of sequence space.

Despite those differences, both the classical Darwinian and the neutralist theories favor the idea that wild-type populations will localize sharply in sequence space after completing an advantageous or neutral shift. Also, both theories assume that mutations appear blindly, irrespective of their selective value. No single neutral or advantageous mutation would occur more frequently than any disadvantageous one.

That view, however, is not sustained by the modern kinetic theory of molecular evolution, nor is it backed by experiments with viruses. After all, evolutionary selection is a consequence of the ability of a genome to replicate itself accurately. Imagine a case in which the process of replication is so highly error-prone that no copy resembled its parental sequence. The resulting population would behave like an ideal gas, expanding until it filled the sequence space at a very low density. Selection acting on such a population could not define it or confine it in any way. The population would lose all its integrity.

If we were to reduce the error rate of replication progressively, variation in the population would disperse less and less as the offspring came to resemble their parents more and more. At some critical error rate, the effect of selection on the population would change radically: the expansive force of mutation would strike a balance with the compressive force of selection. The diffuse gas of related sequences would suddenly condense into a finite but extended region.

This region in sequence space can be visualized as a cloud with a center of gravity at the sequence from which all the mutations arose. It is a self-sustaining population of sequences that reproduce themselves imperfectly but well enough to retain a collective identity over time. Like a real cloud, it need not be symmetric, and its protrusions can reach far from the center because some mutations are more likely than others or may have higher survival values that allow them to produce more offspring. That cloud is a quasispecies.

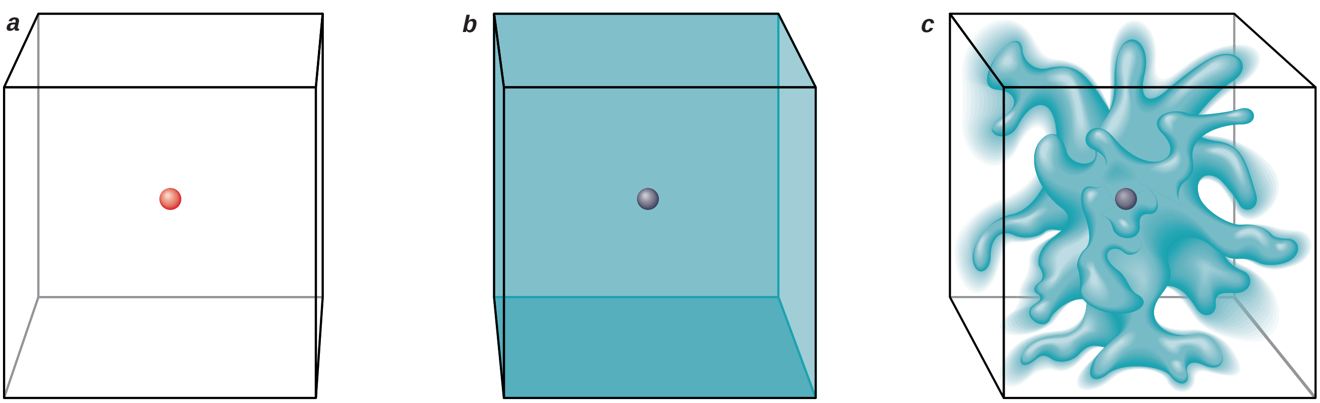
Biologically, the quasispecies is the true target of selection. All the members of a quasispecies—not just the consensus sequence—help to perpetuate the stable population. The fitness of the entire population is what matters, not the fitness of individual members. The wild type of a quasispecies refers to an average for all the members, not to a particularly fit individual. Chemically, the quasispecies is a multitude of distinct but related nucleic acid polymers. Its wild type is the consensus sequence that represents an average for all the mutants, weighted to reflect their individual frequency. Physically, the quasispecies is a localized distribution in sequence space that forms and dissolves cooperatively in very much the same way that molecules of water pass through phase transitions as they freeze or evaporate. Its stability is constrained by the error threshold, which may be interpreted as a kind of "melting point" for the genome information. The population density at each point of sequence space depends on the fitness value of that particular sequence. A mathematician would describe the distribution of sequences in a quasispecies with a vector that refers to the maximum growth within the set of coupled kinetic equations for all the mutants.

One might wonder why in this model an advantageous or neutral mutant would have a better chance to occur than a deleterious one. New mutants appear at the periphery of the quasispecies distribution, where they are produced by the erroneous copying of mutants already present. Because the population of a mutant in the quasispecies depends on its degree of fitness, well-adapted mutants have a better chance of producing offspring; deleterious mutants produce no offspring at all. Because the chance of finding a well-adapted or advantageous mutant is greatest in a region of sequence space associated with high fitness, there is a large bias toward producing such well-adapted mutants. Calculations show that this effect speeds up the evolutionary opportunization of viruses by many orders of magnitude, as compared with truly random, unbiased mutations.

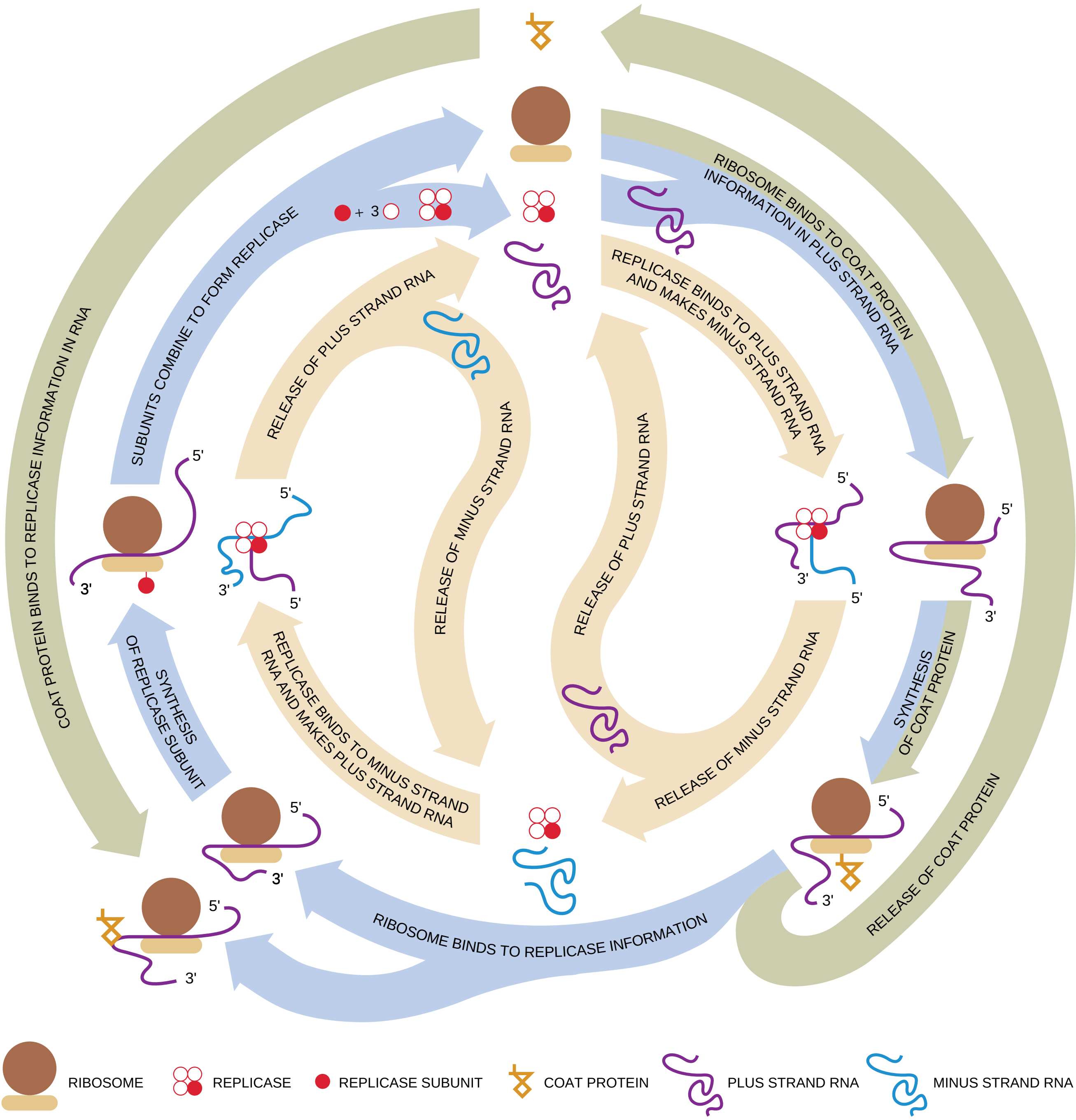
Because the error rate directly determines the size and integrity of a quasispecies, it is the most telling characteristic of a virus. The error rate is the probability that an error will occur when one nucleotide in a sequence is being copied. It can depend both on the type of nucleotide substitution taking place and on its position in the sequence. The position is important because the ribosome interprets the nucleotides three at a time, in a group called a codon. In most codons the first two positions suffice to specify the amino acid to be incorporated into a protein. Mutations in the first two positions may therefore be more stringently maintained by selection. When researchers speak of the error rate of an entire viral sequence, they are referring to an average for all the positions.

In general, the error rate of a virus is roughly proportional to the reciprocal of its sequence length—that is, about one error per replicated sequence. If the error rate were much larger, almost every replication event would produce an unfit mutation. For an entity that produces as many offspring as a virus, an error rate reciprocal to the sequence length is highly significant. Consider a typical infection process, which starts when at least one viable virus enters a host organism. If that virus is not eradicated, it will replicate. Before an infection is detectable, the viral population must rise to around 10^9, which would take about 30 generations. If the error rate is more or less equal to the reciprocal of the sequence length, then on average one error will have been added in each generation.

Consequently, any two viruses taken from an obviously infected host are likely to differ from each other at 30 nucleotide positions or more. When researchers first noticed the sequence diversity of the HIV viruses they found in individual patients, they thought it was evidence of multiple infections by different strains. The work of Simon Wain Hobson of the Pasteur Institute in Par-



POPULATION DYNAMICS of a virus depend on the error rate of its replication process. These figures are highly simplified representations of the sequence spaces that might contain a viral population. If the replication process of a virus were perfectly accurate, all the viral offspring would occupy the same position in sequence space (*a*). If replication were highly imperfect, mutant viruses would soon occupy every position in sequence space (*b*), and the viral population would lose its integrity. At some intermediate error rate, however, the viral population would become a coherent, self-sustaining entity that resembles a cloud centered on the original consensus sequence (*c*). That cloud is a quasispecies.



HYPERCYCLES govern the replication of viruses inside host cells. A hypercycle consists of interlocked feedback loops. In the replication of the plus-strand bacteriophage virus Qß, for example, the reproduction cycle (*tan*) for the genetic information is promoted by a second cycle (*blue*) involving the production of a viral replicase enzyme. At the same time, viral replication is inhibited by the production cycle (*green*) of the viral coat protein, which prevents the synthesis of replicase subunits. The combined influence of these cycles determines the proportions in which viral components are made and thereby the rate of viral replication. Because errors can accumulate in the hypercycle, viruses are prone to mutation.

is has demonstrated, however, that the diverse HIV sequences in patients are usually related to one another. His work clearly confirms that viruses, and immunodeficiency viruses in particular, are quasispecies.

The proliferation of a viral quasispecies is a more complex phenomenon than the simple replication of a wild type. Viral replication takes the form of a hypercycle, a set of interlocking feedback loops that describes a regulated coevolution within a cell of the viral genes and the viral proteins essential to replication that are encoded by those genes. Michael Gebinoga of the Max Planck Institute for Biophysical Chemistry in Göttingen has quantified the process in vivo for the Qß bacteriophage. He found evidence of two feedback cycles, one based on the enzyme replicase, which promotes replication, and the other based on the viral coat protein, which limits it. The first molecules of replicase and other proteins produced by the infectious plus strand are fairly accurate because most copies of the viral genes in the cell are similar to the originals. Errors accumulate mostly during later stages in the infection cycle. For that reason, the synthesis of replicase seems to occur primarily early after infection. Yet even viral sequences that make defective proteins are copied because the replicative machinery acts on all the strands indiscriminately. When an infected *E. coli* cell bursts after 40 minutes, it releases around 10,000 phage particles, of which only 1,000 or less are infectious.

Analyses of sequence space can reveal information about the evolution of viral quasispecies that would otherwise be inaccessible. A straightforward procedure for studying the evolution would be to follow the changes in a viral gene over time. A researcher would need to collect samples of a virus over a period of many successive years. The difficulty is that even for quickly mutating viruses, the amount of change that can accumulate in only a few years—say, the lifetime of a Ph.D. thesis—is too small to measure meaningfully. Hence, the experiment would never be done.

In the mid-1980s Peter Palese of Mount Sinai School of Medicine found a better way. He was lucky enough to obtain samples of influenza A virus that had been isolated and frozen during outbreaks of the disease over a span of about 50 years. Palese and his co-workers analyzed the gene sequence common to those samples. From that information, they plotted the evolutionary relations among the viruses from each epidemic. The "family tree" they created shows the worldwide spread of the virus from a common source in successive waves during each epidemic. The tips of the branches are the isolated virus samples; the nodes, or connections of branches, correspond to the consensus sequences of their shared ancestors. In collaboration with Walter M. Fitch of the University of California at Irvine, Palese found for influenza A an essentially linear relation between the degree of difference for any two sequences and the amount of time since their divergence. Depending on the sequences they examined, two to four mutations appeared per year. The tip-to-node distances on the tree, which reflected the spread of individual sequences, corresponded to roughly five years of evolution.

Unfortunately, the case of influenza A is as yet unique: no other collections of viruses that extend across 50 years currently exist. Nevertheless, other researchers have made progress by employing a different approach. Whereas Palese tracked the evolution of a virus over time, those workers have reconstructed evolutionary trees by making inferences from the similarities of different viruses and viral strains that abound at approximately the same time. Gerald Myers of Los Alamos National Laboratory has made such a tree for the AIDS-causing strain HIV-1, using samples collected from 1985 to 1987.

The principal difference between the tree for HIV-1 and that for influenza A virus is the length of their branches. According to the scheme Myers developed, all the early strains of HIV-1 came from African sources. Looking at the tree, we can almost trace the journey of the virus from that continent to the rest of the world. Indeed, one can extend the tree even further back into evolution by finding the relations between HIV-1, HIV-2 and various forms of simian immunodeficiency viruses (SIVs).

For determining when these viruses diverged, it would be helpful if the separation in the sequences could be used as a measure of evolutionary time. Sadly, the problem is not that simple. If two long, originally identical sequences mutate randomly, it is at first unlikely that they will undergo the same changes at the same positions. Mutations will increase their distance from the original consensus sequence, and those changes will accumulate almost linearly with respect to time.

Eventually, however, when enough mutations have accumulated, some of them will probably reverse a previous change or duplicate a change in the other sequence. As a result, the amount of difference between the sequences will decrease or stay constant, and their distance from the original consensus sequence will finally fluctuate around a given value. Past a certain point, then, the passage of more time does not add more distance. For a genetic sequence in which any one of the four nucleotides could occupy any position, that distance is 75 percent of the total sequence length.

Moreover, the assumption of uniform substitution probabilities is usually not correct. Some positions are almost constant because of fitness constraints; some vary at a normal rate, whereas still others are hypervariable and change rapidly in response to the selection pressure imposed on them by the immune response of their host. The constant, variable and hypervariable positions would each evolve according to a different distance-time relation. Applying different relations to an interpretation of the evolutionary distances would give results for old divergences that differed by orders of magnitude. The lengths of the branches in the evolutionary trees cannot divulge when new viruses evolved.

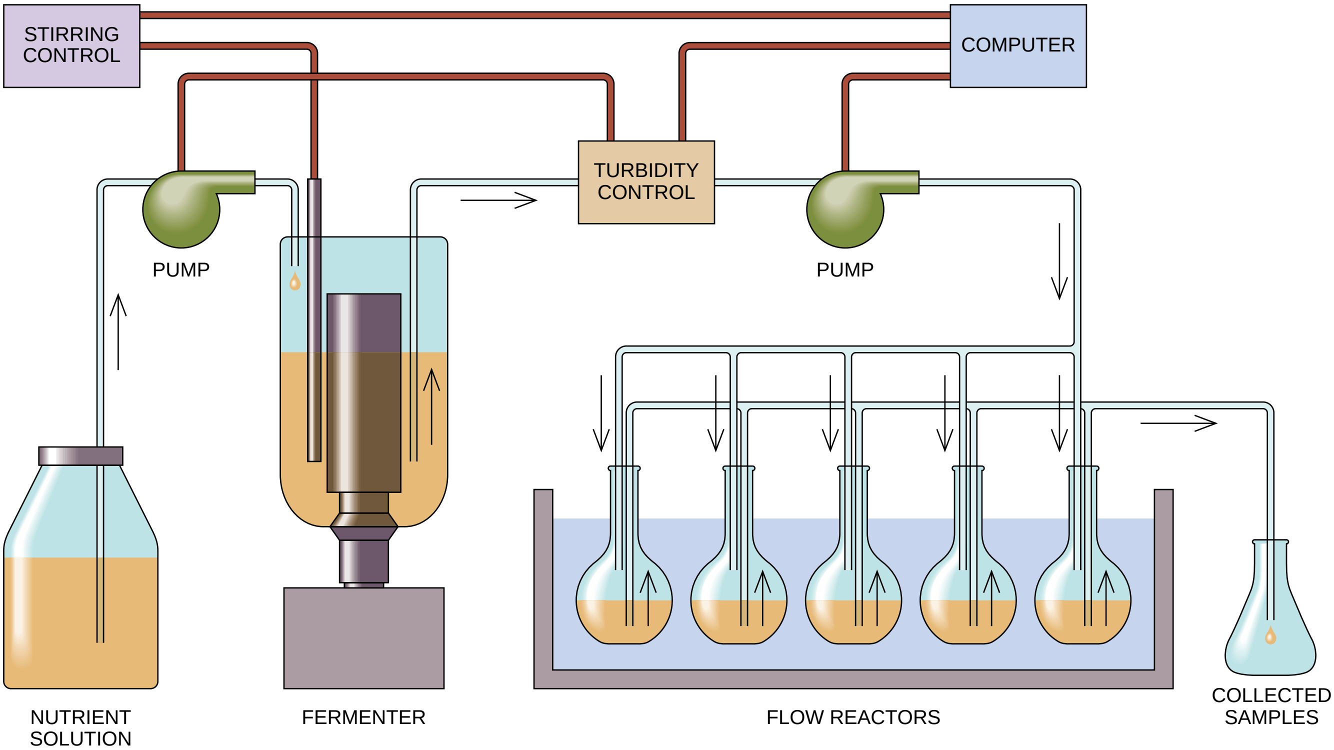
Sequence space diagrams can, however. My colleagues Katja Nieselt-Struwe and Ruthild Winkler-Oswatitsch of Göttingen, Andreas Dress of the mathematics department of Bielefeld University and I have taken that approach. We developed a mathematical method of analyzing the relations within a quasispecies that we call statistical geometry in sequence space. That analysis allows us to determine how often on average different types of changes occur at different positions. It enables us to classify different positions in the viral sequences as constant, variable or hypervariable. From that information, we can deduce roughly how long different viral lineages have existed and the frequency with which different types of mutations occur.

What do the statistical geometries of the influenza A, polio-1 and immunodeficiency viruses reveal? For the tree of influenza A virus, the probability of mutations that would parallel or reverse previous changes is small. As Palese's study indicated, the amount of difference between strains of the virus increases almost linearly over time. An intriguing prediction also emerges from the data: if all the mutable positions in the virus continue to change at the indicated rates, the influenza virus should completely lose its identity within a few hundred years. Because some positions must be constant, the influenza A virus will probably remain a pathogen, because to survive, it will need to infect humans, but we cannot predict what its pathology will be.

For polio-1 virus, the picture is entirely different. In the studied sequence segment, the nucleotides that occupy the first and second positions in each codon scarcely change at all. Mutations at those positions must be strongly eliminated from the quasispecies by selection. Conversely, the nucleotides at the third codon positions are almost completely randomized. As a result, even though the poliovirus has about the same error rate as the influenza virus, only mutations that do not change the encoded amino acids appear in the quasispecies. The proteins in the poliovirus are very highly conserved.

The immunodeficiency viruses have a third type of statistical geometry. All three codon positions are appreciably randomized for all types of changes. We have been able to determine the prevalence of constant, variable and hypervariable sites within the gene for an HIV surface protein that we analyzed. From that information, we were able to estimate how long it must have taken for the immunodeficiency viruses to have diverged to the observed degree.

About 20 percent of the positions are constant, apparently because they are necessary for HIV to function as a retrovirus. They establish that HIV is the



"EVOLUTION MACHINES" of various types are used in the author's laboratory to study the evasive changes that virus populations can make when subjected to selection pressure. The machines create systems of cell cultures in which viruses grow under tightly controlled conditions for many generations. Nutrient solution is pumped into a fermenter in which grow host cells, such as the bacteria *Escherichia coli*. These cells are then pumped into an array of environmentally controlled vessels called flow reactors, where the viruses can parasitize their hosts. Samples of the virus populations can be withdrawn from the flow reactors for analysis. A computer regulates components of the system, such as the pumps and the controls for stirring medium turbidity, that determine the growth conditions and selection pressures on the viruses.

descendant of an old viral family. About 70 percent of the positions are variable and have an average lifetime of about 1,000 years (give or take a few hundred). They seem to give HIV its specific characteristics. Many of these positions differ in HIV-1, HIV-2 and the SIV sequences, which indicates that they must have evolutionarily diverged long ago. My colleagues and I estimate that it was 600 to 1,200 years ago (or even longer, because more constant positions may yet be hidden in the data). Contrary to the evidence of the epidemiologic curves, therefore, HIV is not a new virus, although its pathogenicity may have varied over the centuries.

About 200 positions in the studied HIV gene—about 10 percent of the total—are hypervariable and change on average within 30 years. They provide the tremendous variability that enables HIV to thwart the attempts by its host's immune system to eliminate it. They may also be directly responsible for much of the damage that the virus does to the immune system. According to a theory advanced in 1992 by Robert M. May and Martin A. Novak and their colleagues at the University of Oxford, HIV uses its capacity for variance to outflank the immune response of its host. The number of different sequences that result from mutations at hypervariable sites outruns by far the capacity of the immune system to generate lymphocytes. If HIV can change at all its hypervariable sites in 30 years, it could exhaust the immune system in only a fraction of that time. The virus can produce mutants that evade the immunologic defenses, particularly because its infection targets are the *T* lymphocytes that control the immune response.

Computer simulations carried out by the Oxford group verify those predictions. That theory, based on the quasispecies nature of the virus, also satisfactorily explains the decade-long delay that usually occurs between the initial viral infection and the fatal state of the disease, when the immune system breaks down fairly suddenly. It may take that many years for HIV to exhaust the adaptive resources of the immune system. New experiments will test whether this explanation is correct.

The statistical geometry data also offer insights into ways of fighting HIV and other viruses. The most common way to rid an infected individual of a virus is to stimulate, activate or support the immune system, as a vaccine does. An awareness of the variational flexibility of viruses suggests that three additional strategies must also be explored to improve vaccines. One is to find stable immunologic features in the viral quasispecies against which highly specific monoclonal antibodies could be directed. The second is to create antibodies that can act against a broad spectrum of the likely mutant viruses that would otherwise permit a quasispecies to escape attack. The third is to spot such escape mutants during an early phase of infection and to outmaneuver them with specific agents before they can produce progeny.

The most fruitful approaches may vary with different viruses. For example, the immune system can quickly learn to recognize the almost constant protein features of the poliovirus. That virus has no chance of surviving if it encounters a vaccinated host. The real effectiveness of that protection became apparent only recently when researchers discovered that the mild strain of polio-1 virus in the Sabin vaccine differs from the pathogenic wild type at only two nucleotide positions. It is entirely possible, therefore, that a few of the polioviruses from a vaccine do mutate into a pathogenic state inside the host. Yet by the time those mutations occur, the immunologic protection of the host is already practically perfect. The success of the Sabin vaccine in saving the lives of countless children is unchallenged.

Influenza is a quite different case, as are other viruses. The targets for the immune response against influenza change steadily. Although the immune system eventually copes with the virus and quells the infection, there is no lasting protection. As a consequence, people can contract influenza repeatedly, and new vaccines must be prepared every few years. John J. Holland of the University of California at San Diego and Esteban Domingo of the Independent University of Madrid have observed that the viruses responsible for foot-and-mouth disease and vesicular stomatitis, an infection of the oral membranes in livestock, behave in a similar way. HIV, with its many variable and hypervariable positions, mutates even more rapidly and radically. Vaccines may not have any lasting value against such infections.

VACCINATION has been extremely effective in controlling polio and some other diseases. Because the proteins of poliovirus change very little over time, it is relatively easy to find consistently good immunologic targets. Against more mutable viruses, such as the AIDS virus, vaccination is much less potent.

But vaccines are only one way to fight viruses. The administration of drugs that block viral replication is an extremely common therapy—and for AIDS it is currently the sole therapy that is in any way effective at slowing the progress of the disease. In theory, artificial chains of RNA could be administered to patients to prevent or eliminate viral infections. Those RNA molecules would hinder viral replication, either by binding to the viral RNA or by competing with it for essential enzymes. Specific factors that interfere with viral replication could also be incorporated into host cells by genetic technology. Yet all these approaches may have harmful side effects or would need to clear significant technical hurdles.

A further complication is that viruses may be able to mutate around such obstacles. In my laboratory Björn F. Lindemann has used the understanding of the replicative mechanism of the Qß bacteriophage to test one antiviral strategy. He inserted the gene for the viral coat protein into cells. The cells became resistant to infection because the coat protein, a natural regulator of the phage's replication, blocked the transcription of viral genes.

Yet this strategy did not work perpetually: given sufficient time and generations, the Qß bacteriophage adapted by mutating into a form that ignored the coat protein signal. Lindemann demonstrated that fact using one of the automated "evolution machines" developed recently in my laboratory. In these devices, viruses grow in host cells for extended periods under mild selection pressures. Evolutionary biotechnology, or applied molecular evolution, as it is often called, is a rapidly emerging field of research that may have many applications in new antiviral strategies [see "Directed Molecular Evolution," by Gerald F. Joyce; SCIENTIFIC AMERICAN, December 1992].

One strategy may be resistant to the evasive maneuvers of viruses: it would exploit their nature as quasispecies and thereby undermine the very basis of their existence. Even in a successful viral quasispecies, only a small fraction of the viral sequences in a host cell are viable. If the error rates of viruses can be increased moderately, just enough to cross the critical error threshold that defines their quasispecies, they would experience a catastrophic loss of information. The viral quasispecies would fall apart because it would be producing too many nonviable mutants.

Using drugs that produce mutations, Domingo and Holland have demonstrated that this approach works against the virus that causes foot-and-mouth disease. For such a strategy to work as a therapy, however, the drugs must change the error rate of only the viral replicase and not of enzymes essential to the host's well-being. Careful study of replicase mechanisms should bring about such a possibility of interfering with virus infection. This strategy would be precisely the opposite of immunization therapies that attempt to prevent the appearance of escape mutants.

As of today, we know little about the origin of viruses or their role in the evolution of the biosphere. Viruses come and go: some adapt; others disappear. The undeniable reality is that an estimated 13 million people worldwide are infected with HIV. Pandora's box is still open and releasing new ills. Nevertheless, our growing understanding of viruses suggests that, as in the original myth, hope has not escaped.

FURTHER READING

MOLECULAR QUASI-SPECIES. Manfred Eigen, John McCaskill and Peter Schuster in *Journal of Physical Chemistry,* Vol. 92, No. 24, pages 6881–6891; December 1, 1988.

ROLE OF GENOME VARIATION IN VIRUS EVOLUTION. Manfred Eigen and Christof K. Biebricher in *RNA Genetics,* Vol. 3: *Variability of RNA Genomes.* Edited by Esteban Domingo, John J. Holland and Paul Ahlquist. CRC Press, 1988.

HOW OLD IS THE IMMUNODEFICIENCY VIRUS? Manfred Eigen and Katja Nieselt-Struwe in *AIDS,* Vol. 4, Supplement 1, pages S85–S93; 1990.

STATISTICAL GEOMETRY ON SEQUENCE SPACE. Manfred Eigen and Ruthild Winkler-Oswatitsch in *Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences.* Edited by Russell F. Doolittle. Academic Press, 1990.

THE HYPERCYCLE: COUPLING OF RNA AND PROTEIN BIOSYNTHESIS IN THE INFECTION CYCLE OF AN RNA BACTERIOPHAGE. M. Eigen, C. K. Biebricher, M. Gebinoga and W. C. Gardiner, Jr., in *Biochemistry,* Vol. 30, No. 46, pages 11005–11018; November 19, 1991.

Australia's Polar Dinosaurs

Their excellent night vision and apparent warm blood raise a question: Could they have survived icehouse conditions at the end of the Cretaceous period?

by Patricia Vickers-Rich and Thomas Hewitt Rich

In the Early Cretaceous period, about 100 million years ago, Australia lay alongside Antarctica, which straddled the South Pole as it does today. Australia's southeastern corner, now the state of Victoria, lay well inside the Antarctic Circle. At that time, the region hosted an assemblage of animals and plants that lived under climate conditions having no modern analogue. The average temperature, though low, appears to have been within the temperate range, yet the sun did not shine throughout the long winter.

Many dinosaur lineages survived in this strange environment after they had died out in other places. At least one member of the group evolved an adaptation to the cold and to the dark that is interesting both in itself and for what it tells of the passing of a biological epoch. If global cooling indeed killed the dinosaurs, as many paleontologists have suggested, then Australia's species were the ones most likely to have survived the longest. Did their adaptations to an already marginal climate help them survive a sharp cooling trend, one that caught species living on other continents unprepared?

Although the Cretaceous fossil plants of southeastern Australia have been studied for more than a century, the

LEAELLYNASAURA *ALLOSAURUS* *MUTTABURRASAURUS*

AUSTRALIAN DINOSAURS flourished in southeastern Victoria during the Early Cretaceous, when the region lay within the Antarctic Circle. This mural depicts six species that left fossils there and a seventh—the large iguanodontid *Muttabur-*

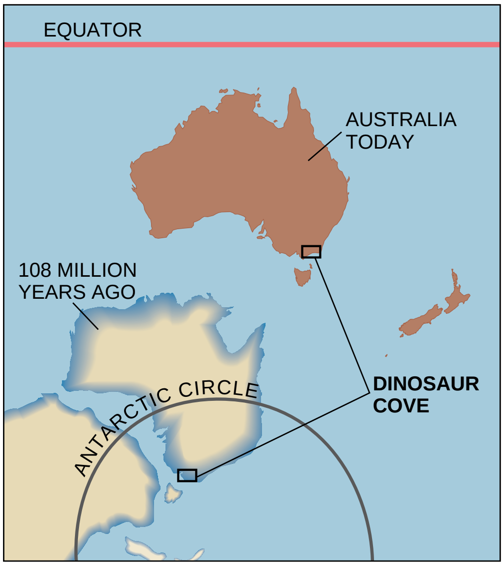
animals remained mostly hidden until recently. Around 1900 the geologist William Hamilton Ferguson found two bones that have had a bearing on later paleontological work—the tooth of a lungfish and the claw of a carnivorous dinosaur, assigned to the theropod genus *Megalosaurus.* For the next 70 years, as no further finds joined them, these bones lay neglected in a cabinet in the Museum of Victoria. Then, in 1978, two graduate students at Monash University, Tim F. Flannery and John A. Long, discovered near Ferguson's original site the first specimens of a trove of dinosaur bones embedded in hard sandstones and mudstones from the Early Cretaceous.

These discoveries—only an hour and a half's drive southeast of Melbourne—encouraged paleontologists to prospect other coastal sites. In 1980 we struck a rich lode in the Otway ranges, which the Victorian government, at our suggestion, has since named Dinosaur Cove. There, with the help of Earthwatch and other volunteers and the National Geographic Society, the Australian Research Council and Atlas Copco, a manufacturer of mining equipment, we have spent three months out of every year chiseling, hammering and on occasion blasting tunnels into the fossil-bearing strata [*see illustration on page 55*]. This project has been the center of our lives and the lives of our co-workers, our children and even our parents (two of whom are paleontologists).

Dinosaur Cove and other sites of similar character were formed when violent, seasonal streams swept broad floodplains of their accumulated bones and plant life, depositing this flotsam and jetsam at the bottom of shallow stream channels. These deposits appear along the southern Victorian shore because only there could gnawing waves expose the sediments laid down in the rift valley that formed when Australia and Antarctica went their separate ways, as did the other fragments of Gondwana, the ancient supercontinent [*see illustration on next page*]. Only two fossil sites from the same period have been found inland, one in sediments laid down under far quieter conditions at the bottom of an

PATRICIA VICKERS-RICH and THOMAS HEWITT RICH collaborate in the study of fossils. Vickers-Rich is a reader in earth sciences and in ecology and evolutionary biology at Monash University in Melbourne, Australia. She is interested in reconstructing ancient environments, especially those without modern analogues, and in analyzing rapid biotic change, such as mass extinctions. Rich is curator of vertebrate paleontology at the Museum of Victoria in Melbourne. He conducts research on the evolutionary patterns of Mesozoic vertebrates, specializing in primitive mammals and ornithischian dinosaurs. The Riches received undergraduate degrees in paleontology from the University of California at Berkeley and doctorates in geology from Columbia University. They live near Melbourne and have two children.

***rasaurus*—that has been found only in Queensland, far to the north. The paucity of large polar dinosaurs may reflect a real absence or merely the selective preservation of small bones. Peter Trusler painted the mural for Australia Post.**



SOUTHERN SUPERCONTINENT began to break up more than 100 million years ago, when a rift valley formed between Australia and Antarctica (*left*). Stream channels in the valley received bones gathered by floodwaters that periodically swept these broad plains. The bones, together with clay and silt, created the fossil-bearing formations of Dinosaur Cove (*right*).

ancient lake. This inland site has therefore yielded some uncommonly well preserved specimens.

It must be noted that southeastern Australia's dinosaurs are known from a mere 5,000 individual bones and two partial skeletons. Just a few hundred of the bones can be assigned to a given species or genus. What they lack in number, however, they make up in scientific interest.

All efforts at interpretation revolve around the estimation of temperature, for which two methods have been tried. Robert T. Gregory of Southern Methodist University and his associates infer Australian paleoclimate from the ratio of oxygen 18 to oxygen 16 trapped in ancient rocks. They find that mean annual temperatures probably approached zero degrees Celsius but might have reached as high as eight degrees C above zero. Such values occur today in Hudson Bay, Saskatchewan (zero degrees C), and in Minneapolis and Toronto (eight degrees C above zero).

Robert A. Spicer of the University of Oxford and Judith Totman Parrish of the University of Arizona instead deduce temperature from the structure of ancient plants, arriving at the somewhat higher mean annual temperature of 10 degrees C. Their research has demonstrated that polar Australia supported conifers, ginkgoes, ferns, cycads, bryophytes and horsetails but only a few angiosperms, or flowering plants, identifiable by a sprinkling of pollen. The angiosperms were then just beginning to spread into new niches. Perhaps they got their start by exploiting weedy ecological systems in the rift valleys that formed as the supercontinent split apart.

Evergreens, which provided forage in all seasons, had thick cuticles and other structural features that indicate adaptation to cold or dryness (perhaps brought on by winter freezing). Deciduous plants offer another climatic clue: they seem to have lost all their leaves at once. These mass falls may have been triggered by darkness or cold. Drought, however, probably did not serve as a constant cue—the sedimentary record and the abundance of ferns and bryophytes argue for conditions that were moist in all seasons except perhaps winter.

If the higher estimate of mean temperature is correct, Australia was both temperate and subject to a period of continuous darkness every year—a combination with absolutely no modern counterpart. The winter night lasted between six weeks and four and a half months, depending on the true paleolatitude. Because the lower extreme of temperature would then have fallen well below the mean, most of the vertebrates preserved as fossils must have lived quite close to their thermal limits. Some, such as lungfish, cannot now breed in waters colder than 10 degrees C.

If, on the other hand, the lower mean temperature is correct, it becomes more than a normal scientific challenge to understand how this paleocommunity functioned at all. Before seriously attacking this problem, scientists will first have to demonstrate that it exists. To refine the estimate of the average annual temperature, a multidisciplinary team is comparing floral, geochemical and other forms of evidence.

Nothing in this fauna is quite so peculiar to the region as the koala is today, for although the species and genera were local, they belonged to cosmopolitan families. Yet their adaptations are striking, as is the fact that some survived beyond the time of demise for their families elsewhere.

Among such anachronisms—or relicts—are the labyrinthodont amphibians, ancestors of modern amphibians and reptiles. Most paleontologists had thought this group went extinct by the Jurassic, some 160 million years ago. In the past 15 years, however, Michael Cleeland and Lesley Kool of Monash University found three jaws from this group in Victorian sediments dating from the Early Cretaceous. Two of the jaws were unmistakable, because their teeth had the labyrinthine infolding of the enamel that gives this group its name. At least one large species of labyrinthodonts lived in polar Australia 115 million years ago, several million years after the group had died out elsewhere.

How did they survive? We suspect that the cool weather preserved the animals from competition with crocodiles, which were probably poorly adapted to the conditions prevailing in southeastern Australia until the onset of climatic warming during the last five million years of the Early Cretaceous. The hypothesis rests on the fact that contemporary crocodilians now live in waters no colder than 10 degrees C, whereas some modern frogs and salamanders can be active in meltwater from snow.

Another late survivor was the familiar *Allosaurus,* a carnivorous theropod. Elsewhere in the world this animal ranged up to five meters in height, but the southeastern Australian specimen stood no more than two meters high—hardly taller than a human. This "pygmy" is the latest-surviving allosaur that has yet been found. It remains unclear whether this species also owed its longevity to some niche that cold climate may have carved out for it. The discovery of juvenile forms (but no eggshells, as yet) does suggest that these dinosaurs were not just casual visitors but lived near the pole for much of the year, using the area as a nursery during the period of maximum sunlight.

Unlike the allosaurs, many dinosaurs of Australia were not the last in their lineage; some may have been the first. At least two and perhaps as many as four families of dinosaurs have been recognized that include forms which are either the oldest or among the oldest of their kind. For instance, the ornithomimosaurs, carnivores of ostrichlike size and appearance, are manifestly primitive and among the oldest within this group; only a Late Jurassic species from East Africa predates the Australian form. The elongated, slender hind limbs of the Australian species made them the gazelles of the dinosaur world, able to escape from predators and to run down prey. The ornithomimosaurs probably originated in Gondwana and spread northward to join the later Cretaceous faunas of North America and Eurasia, where they enjoyed wide success.

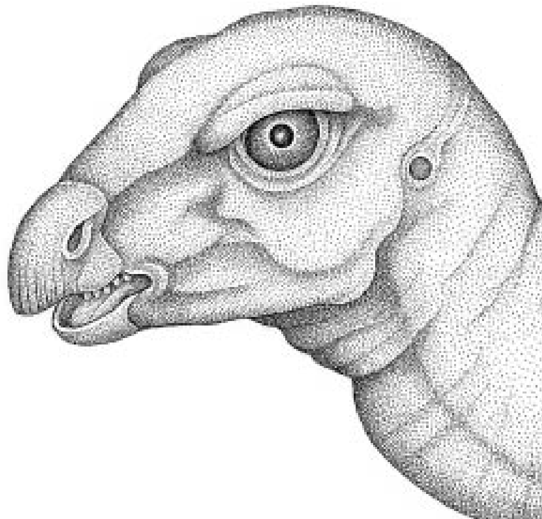
Two very small theropods remain unidentified, but one seems to resemble an egg-eating oviraptosaur, known until now exclusively from the younger Cretaceous rocks of North America and Asia. These groups may also have an origin in Gondwana.

Yet another dinosaur group that has recently been identified belongs to the neoceratopsians, or horned dinosaurs. Identification is tentative, as it is based on just two ulnae (part of the lower arm), but the similarity to *Leptoceratops,* a browser the size of a sheep, is uncanny. Previously, all neoceratopsian records dated from the Late Cretaceous and, with the exception of a few bones from Argentina, came from the Northern Hemisphere. This dinosaur family may also have arisen in the southern supercontinent.

The Early Australian Cretaceous also reshaped forms that continued to flourish in other regions. By far the most successful such group consisted of the hypsilophodontid dinosaurs. These animals, most of them hardly larger than a chicken, were bipeds built for speed, with large hind legs, small but well-developed hands, substantial tails and—for the most part—herbivorous habits. They thus resembled wallabies in both shape and ecological role.

The family Hypsilophodontidae was common throughout the world from the Middle Jurassic to Late Cretaceous times, but its prominence reaches an absolute and relative peak in the Victorian sediments. Not only do hypsilophodontids constitute most of the dinosaur remains, they are also represented by four to five genera, depending on the taxonomic criteria one uses, and five to six species. Other areas, some much more richly endowed with dinosaur species, never harbored more than three kinds of hypsilophodontids at a time. Something clearly favored the diversification of this group in polar Australia.

A particularly intriguing adaptation of at least one species of polar hypsilophodontid is suggested by the magnificently preserved brain cast of *Leaellynasaura amicagraphica* (named after our daughter, friends of the Museum of Victoria and the Nation-



ACUTE NIGHT VISION is suggested by the eyes and brain of *Leaellynasaura amicagraphica,* a hypsilophodontid shown here at life size (*above, right*). The large eyes were common to all hypsilophodontids and may have helped the group dominate an environment marked by seasonal darkness. This hypothesis may also explain the huge optic lobes, of which the left one can be seen at the rear of this natural brain cast (*below; bump at far right*), formed when silt solidified in the skull.

al Geographic Society). The brain, unusually large for a dinosaur of this size, bears the marks of optic lobes whose relative size is easily the greatest ever documented in a hypsilophodontid.

How is one to interpret these enlarged lobes? We hypothesize that they enhanced the animals' ability to see in the dark, enabling them to forage effectively during the long winter months. There would have been no lack of food then, for those capable of seeing it: the herbivores could have lived off evergreens and deciduous leaf mats, and the carnivores could have hunted the herbivores.

This hypothesis also explains why this group came to dominate the polar environment in the first place. Hypsilophodontids everywhere in the world had large eyes and, presumably, acute vision. That trait could have given them their foothold in polar Australia. Once established in this "protected" environment, the hypsilophodontids could have competed with one another to produce the observed diversity of genera and species, perhaps all sharing hypertrophied optic lobes.

If the animals foraged at night, they must have been active at freezing or subfreezing temperatures. This feat goes far beyond the cold tolerance of any modern reptile, even the New Zealand tuatara, *Sphenodon punctatus,* which can remain active at five degrees C provided it can sun itself. *Leaellynasaura* could have survived solely by maintaining a constant body temperature, eating frequently, as birds do in wintertime.

Pterosaurs, flying reptiles, and ankylosaurs, heavily armored dinosaurs, also appear, but in such fragmentary remains as to tell the student little of the animals' lives. Much can be gleaned from one handful of teeth, however, for they come from plesiosaurs. These long-necked reptiles, not themselves dinosaurs, generally paddled the seas, but here they inhabited fresh water in the ancient valley between Australia and Antarctica. They thus recall the Ganges River dolphin, one of the few cetaceans that live in fresh water.

The sauropods alone are absent. These giants, familiar from the example of *Apatosaurus* (or *Brontosaurus,* as it is more popularly known), lived at that time in Australia's lower latitudes. None, however, has been found further south nor, indeed, in any of the nine Cretaceous polar dinosaur sites so far identified in both hemispheres. The only polar sauropod yet discovered is the much older (early Jurassic) *Rhoetosaurus* from northeastern Australia.

The apparent restriction of these large

BONE TO STONE: *Leaellynasaura* as it might have appeared in the process of becoming a fossil. A bone assemblage from an individual could have fossilized in this way only if the stream channel was choked off, forming an oxbow or billabong.

HARD ROCK makes hard work for these volunteer paleontologists. Full-scale mining techniques (*left*) and explosives (*right*) extract fossil-bearing slabs, which tend to fracture along the planes containing the largest treasures.

dinosaurs to lower latitudes in the Cretaceous of Australia may be real or merely an artifact of sampling. We worry about this question because the floodwaters that broke out of rain-swollen rivers would have collected small and medium-size bones but left large ones. The body of a sauropod would have stayed put rather than floating to a place where many specimens were concentrated in the small flood channels, which were no more than five to 10 meters in width and 20 to 30 centimeters in depth.

Yet we suspect there was an underlying tendency toward small body size in these polar environs. None of the hypsilophodontids, it must be remembered, stood taller than a human, and most were barely knee-high. The dwarf *Allosaurus* matches the smallest we have examined in the North American collections. The ornithomimosaur is equally unprepossessing, and the protoceratopsid and the ankylosaur are no bigger than a sheep. A single fragment of a claw constitutes our sole record of a form—a carnivore, apparently similar to *Baryonyx* of England—which may have measured up to eight meters in length.

This pattern contradicts the scaling laws that Bergmann and Allen formulated in the 19th century. According to these laws, animals in a given lineage tend to become larger and more compact as the average temperature of their environment falls. This trend is exemplified by the comparison of mountain lions in Canada with pumas of Central America and of human populations in the subarctic and tropical zones.

Other factors also determine body dimensions, especially the size of the territory in which a population lives. Individuals found on islands are often smaller than their mainland counterparts. For example, there were dwarf elephants on the ancient Mediterranean islands, and pygmy mammoths were recently found in 4,000-year-old sediments on islands off the north coast of Siberia. Dwarfism may be a response to selective pressure to increase the number of individuals so as to ensure a gene pool diverse enough for the species to survive in a restricted area. This effect has also been noted on peninsulas—and ancient southeast Australia was a peninsula of the Gondwana landmass.

The dinosaurs on that peninsula were trapped virtually at the end of the earth. Their direct path north was blocked by a vast inland sea, which they could have passed only by going hundreds of kilometers to the west before wheeling about to the north. At the end of such labors, they would have been able to catch, at most, an hour of sun a day in winter. Migration would have made little sense for such small animals.

Less formidable barriers sealed in the dinosaurs of the one other polar site that has yielded large quantities of fossils: the north slope of Alaska. The dinosaurs there had a clear north-south corridor along which they could migrate with ease. It is significant that those dinosaurs were big—at least equal in size to caribou, wildebeest and other modern animals that migrate.

One must question whether animals so superbly adapted to the cold and the dark could have been driven to extinction by an artificial winter, such as is supposed to have followed a cataclysmic event at the boundary between the Cretaceous and Tertiary formations. It is proposed that the cataclysm, perhaps a collision with a comet or asteroid or a series of volcanic eruptions, suffused the atmosphere with a blanket of dust, excluding sunlight and freezing or starving most animals to death.

We suspect, however, that no such artificial winter could have killed the dinosaurs unless it lasted for a long time, certainly more than a few months. Otherwise at least a few of the polar dinosaurs would have survived the cataclysm. Of course, it is possible that some other development had already ended the reign of southern Australia's dinosaurs by the end of the Cretaceous.

Arthur Conan Doyle once dreamed of a plateau in South America that time forgot, where dinosaurs continued to reign. Reports earlier this year that dwarf mammoths survived to early historical times, in islands off the coast of Siberia, give force to such speculation. If dinosaurs found a similar haven in which they outlived the rest of their kind, then we think polar Gondwana, including southeastern Australia, is a likely place to look for it.

FURTHER READING

EVIDENCE FOR LOW TEMPERATURES AND BIOLOGIC DIVERSITY IN CRETACEOUS HIGH LATITUDES OF AUSTRALIA. P. V. Rich, T. H. Rich, B. E. Wagstaff et al. in *Science,* Vol. 242, pages 1403–1406; December 9, 1988.

OXYGEN ISOTOPIC COMPOSITION OF CARBONATE CONCRETIONS FROM THE LOWER CRETACEOUS OF VICTORIA, AUSTRALIA. R. T. Gregory, C. B. Douthitt, I. R. Duddy et al. in *Earth and Planetary Science Letters,* Vol. 92, No. 1, pages 27–42; February 1989.

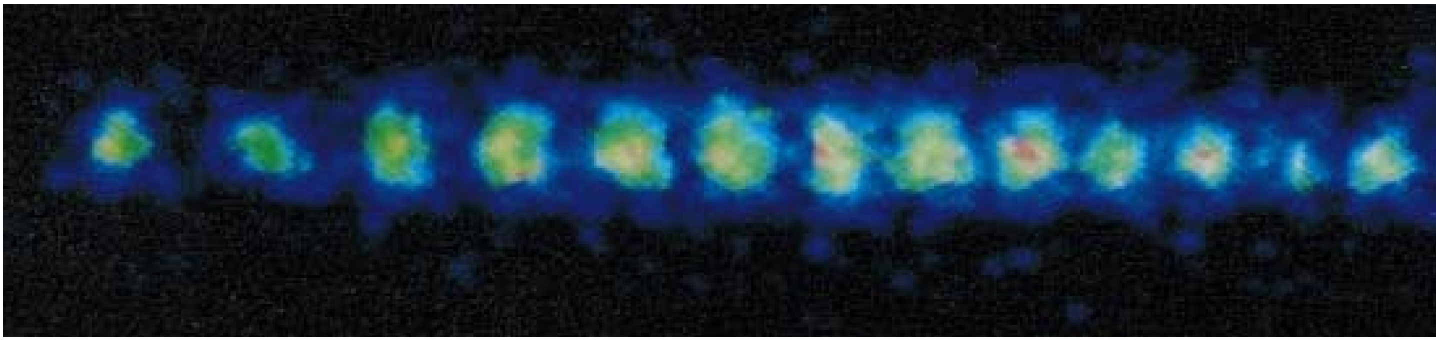
POLAR DINOSAURS AND BIOTAS OF THE EARLY CRETACEOUS OF SOUTHEASTERN AUSTRALIA. T. H. Rich and P. V. Rich in *National Geographic Research,* Vol. 5, No. 1, pages 15–53; Winter 1989.

CONTINENTAL CLIMATE NEAR THE ALBIAN SOUTH POLE AND COMPARISON WITH CLIMATE NEAR THE NORTH POLE. J. T. Parrish, R. A. Spicer, J. G. Douglas et al. in *Geological Society of America, Abstracts with Programs,* Vol. 23, No. 5, page A302; Annual Meeting, 1991.

Accurate Measurement of Time

Increasingly accurate clocks—now losing no more than a second over millions of years—are leading to such advances as refined tests of relativity and improved navigation systems

by Wayne M. Itano and Norman F. Ramsey



Few people complain about the accuracy of modern clocks, even if they appear to run more quickly than the harried among us would like. The common and inexpensive quartz-crystal watches lose or gain about a second a week—making them more than sufficient for everyday living. Even a spring-wound watch can get us to the church on time. More rigorous applications, such as communications with interplanetary spacecraft or the tracking of ships and airplanes from satellites, rely on atomic clocks, which lose no more than a second over one million years.

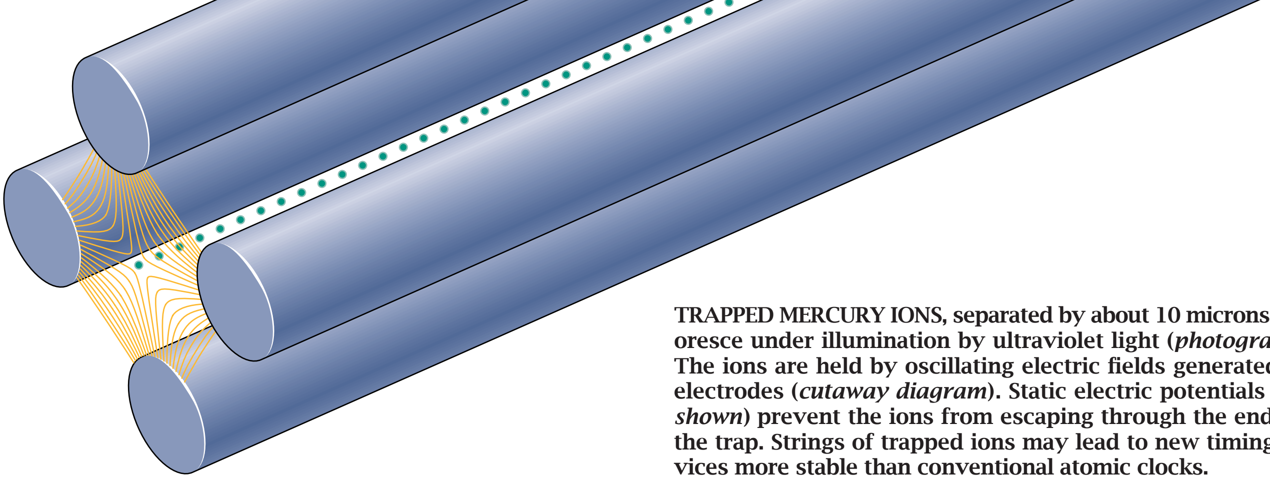
WAYNE M. ITANO and NORMAN F. RAMSEY have collaborated many times before writing this article: Itano earned his Ph.D. at Harvard University under the direction of Ramsey. Itano, a physicist at the Time and Frequency Division of the National Institute of Standards and Technology in Boulder, Colo., concentrates on the laser trapping and cooling of ions and conducts novel experiments in quantum mechanics. He is also an amateur paleontologist and fossil collector. Ramsey, a professor of physics at Harvard, earned his Ph.D. from Columbia University. He has also received degrees from the University of Oxford and the University of Cambridge, as well as several honorary degrees. A recipient of numerous awards and prizes, Ramsey achieved the highest honor in 1989, when he shared the Nobel Prize in Physics for his work on the separated oscillatory field method and on the atomic hydrogen maser.

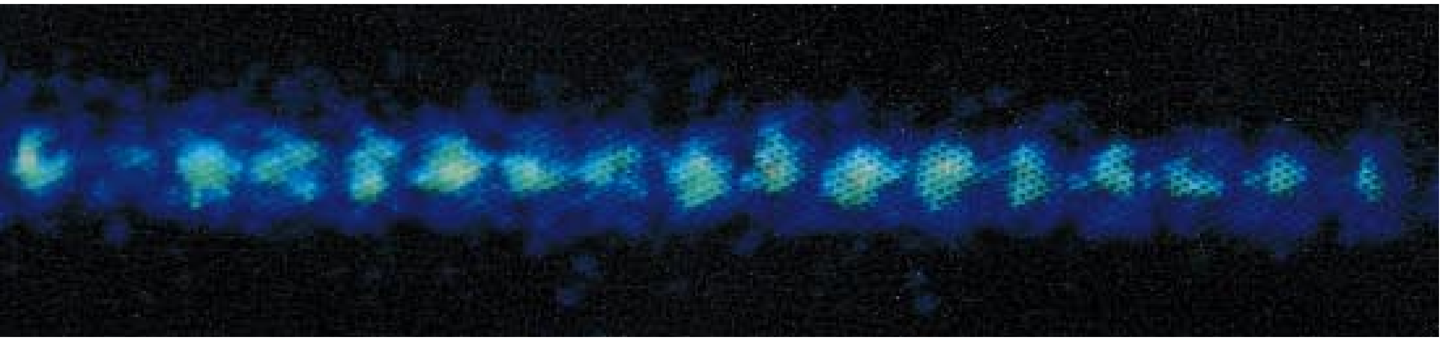
There might not seem to be much room for the improvement of clocks or even a need for more accurate ones. Yet many applications in science and technology demand all the precision that the best clocks can muster, and sometimes more. For instance, some pulsars (stars that emit electromagnetic radiation in periodic bursts) may in certain respects be more stable than current clocks. Such objects may not be accurately timed. Meticulous tests of relativity and other fundamental concepts may need even more accurate clocks. Such clocks will probably become available. New technologies, relying on the trapping and cooling of atoms and ions, offer every reason to believe that clocks can be 1,000 times more precise than existing ones. If history is any guide, these future clocks may show that what is thought to be constant and immutable may on finer scales be dynamic and changing. The sundials, water clocks and pendulum clocks of the past, for example, were sufficiently accurate to divide the day into hours, minutes and seconds, but they could not detect the variations in the earth's rotation and revolution.

A clock's accuracy depends on the regularity of some kind of periodic motion. A grandfather clock relies on the sweeping oscillation of its pendulum. The arm is coupled to a device called an escapement, which strikes the teeth of a gear in such a way that the gear moves in only one direction. This gear, usually through a series of additional gears, transfers the motion to the hands of the clock. Efforts to improve clocks are directed for the most part toward finding systems in which the oscillations are highly stable.

The three most important gauges of frequency standards are stability, reproducibility and accuracy. Stability is a measure of how well the frequency remains constant. It depends on the length of an observed interval. The change in frequency of a given standard might be a mere one part per 100 billion from one second to the next, but it may be larger—say, one part per 10 billion—from one year to the next. Reproducibility refers to the ability of independent devices of the same design to produce the same value. Accuracy is a measure of the degree to which the clock replicates a defined interval of time, such as one second.

Until the early 20th century, the most accurate clocks were based on the regularity of pendulum motions. Galileo had noted this property of the pendulum after he observed how the period of oscillation was approximately independent of the amplitude. In other words, a pendulum completes one cycle in about the same amount of time, no matter how big each sweep is. Pendulum clocks became possible only after the mid-1600s, when the Dutch scientist Christiaan Huygens invented an escapement to keep the pendulum swinging. Later chronometers used the oscillations of balance wheels attached





TRAPPED MERCURY IONS, separated by about 10 microns, fluoresce under illumination by ultraviolet light (*photograph*). The ions are held by oscillating electric fields generated by electrodes (*cutaway diagram*). Static electric potentials (*not shown*) prevent the ions from escaping through the ends of the trap. Strings of trapped ions may lead to new timing devices more stable than conventional atomic clocks.

to springs. These devices had the advantage of being portable.

Considerable ingenuity went into improving the precision of pendulum and balance-wheel clocks. Clockmakers would compensate for temperature changes by combining materials with different rates of thermal expansion. A more radical approach came in the 1920s, when William H. Shortt, a British engineer, devised a clock in which a "slave pendulum" was synchronized to a "free pendulum." The free pendulum oscillates in a low-pressure environment and does not have to operate any clock mechanism. Instead it actuates an electrical switch that helps to keep the slave pendulum synchronized. As a result, the period of the Shortt clock is extremely stable. These clocks had an error of a few seconds in a year (about one part per 10 million) and became the reference used in laboratories.

The next major advance in timekeeping was based on the development of quartz-crystal electronic oscillators. The frequency of such devices depends on the period of the elastic vibration of a carefully cut quartz crystal. The vibrations are electronically maintained through a property of such crystals called piezoelectricity. A mechanical strain on the crystal produces a low electric voltage; inversely, a voltage induces a small strain.

The quartz vibrates at a frequency that depends on the shape and dimensions of the crystal. In some wristwatches, it is cut into the shape of a tuning fork a few millimeters long. In other timepieces, it is a flat wafer. The quartz is connected to an electric circuit that produces an alternating current. The electrical feedback from the quartz causes the frequency of the circuit to match the frequency at which the crystal naturally vibrates (usually 32,768 hertz). The alternating current from the circuit goes to a frequency divider, a digital electronic device that generates one output pulse for a fixed number of input pulses. The divider also actuates either a mechanical or digital electronic display.

In the late 1920s Joseph W. Horton and Warren A. Marrison, then at Bell Laboratories, made the first clock based on a quartz-crystal oscillator. In the 1940s quartz-crystal clocks replaced Shortt pendulum clocks as primary laboratory standards. These clocks were stable to about 0.1 millisecond per day (about one part per billion). Relatively inexpensive, quartz clocks continue to be extensively used. The timekeeping elements of common quartz watches and clocks are simplified and miniaturized versions of quartz frequency standards. Quartz wristwatches became common once the ability emerged to cut the quartz into thin, tuning-fork shapes reliably and to manufacture miniature, low-power digital electronic components.

Yet quartz-crystal clocks prove inadequate for many scientific applications, such as tests of relativity. According to Albert Einstein's calculations, gravity distorts both space and time. The difference in gravitational potential causes time to pass more quickly high in the atmosphere than it does on the surface. The difference is slight. Time runs about 30 millionths of a second per year faster at the top of Mount Everest than it does at sea level. Only atomic frequency standards achieve the requisite precision.

The quantized energy levels in atoms and molecules provide the physical basis for atomic frequency standards. The laws of quantum mechanics dictate that the energies of a bound system, such as an atom, have certain discrete values. An electromagnetic field can boost an atom from one energy level to a higher one. The process can also work in reverse. If the atom is in a high energy level, it can drop to a lower level by emitting electromagnetic energy.

The maximum amount of energy is absorbed or emitted at a definite frequency—the resonance frequency, or the difference between the two energy levels divided by Planck's constant. This value is sometimes called the Bohr frequency. Such frequencies make ideal time standards because they are extremely stable. Time can be kept by observing the frequencies at which electromagnetic energy is emitted or absorbed by the atoms. In essence, the atom serves as the master pendulum whose oscillations are counted to mark the passage of time.

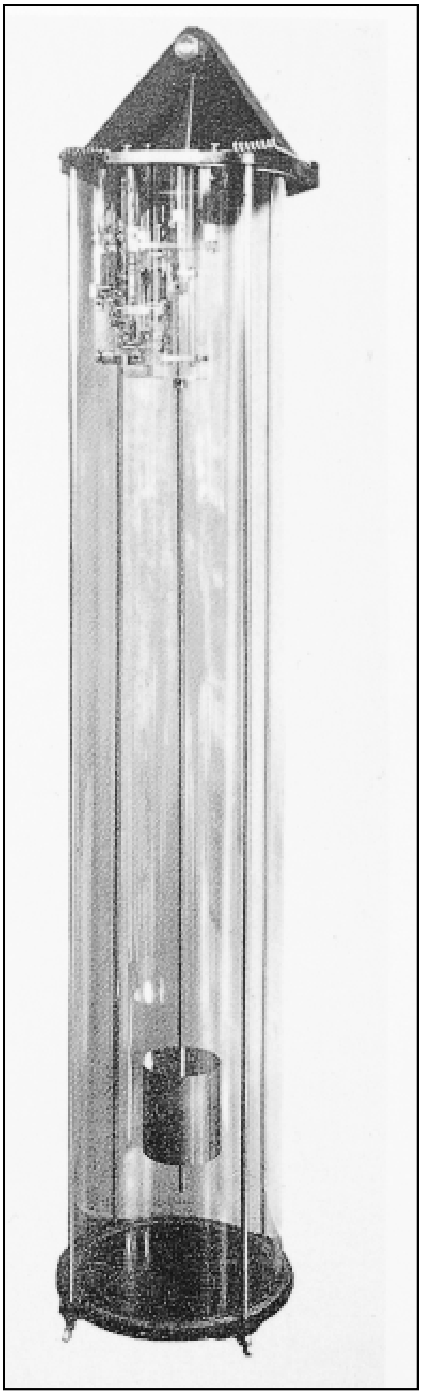
Although we have described general quantum properties, the effects exploited in atomic clocks are slightly more complicated. In most atomic clocks the energy that atoms absorb or release actually results from transitions between so-called hyperfine energy levels. These levels exist because of an intrinsic property of particles known as the magnetic moment. Electrons and the nuclei of most atoms spin about their axes as if they were tops. In addition, they are magnetized, like compass needles oriented along their axes of rotation. These axes can have different orientations with respect to one another, and the energies of the orientations may differ. These positions correspond to the hyperfine levels. The nomenclature comes about because the levels were first observed in spectroscopy as small splittings of spectral lines.

On paper, standards based on atomic processes are ideal. In practice, perfection is elusive. Atoms do not absorb or emit energy precisely at the resonance frequency. Some energy is spread over a small interval surrounding the frequency—a smearing of frequencies, so to speak. All else being equal, the precision to which the resonance frequency can be measured is inversely proportional to this smearing. The greater the spread, the less precise the measurement. The spread is often expressed in terms of the quality factor, or *Q*, which is equal to the resonance frequency divided by the frequency spread. In many cases, the higher the resonance frequency, the higher the *Q*. Furthermore, smearing is often inversely proportional to the time the atom is in the apparatus. In those situations, the *Q* of the resonance, and hence the precision of the measurement, increases as the measuring time increases.

The motions of the atoms also introduce uncertainty by causing apparent shifts in the resonance frequencies. Such changes appear because of the Doppler effect. The phenomenon can be divided into first- and second-order shifts if the atoms are moving much slower than the speed of light. The first-order Doppler shift is an apparent change in the frequency of the applied electromagnetic wave as seen by a moving atom. The amount of the shift is proportional to the velocity of the atom. If the atom moves in the same direction as the wave does, the shift is to a lower frequency. If the atom's motion is opposed to that of the wave, the shift is to a higher frequency. If the directions are perpendicular, the first-order shift is zero.

The second-order Doppler shift comes about as a consequence of time dilation. According to relativity, time slows down for objects in motion; a moving atom "sees" a slightly different frequency than does a stationary counterpart. The effect on the resonance frequency is usually much smaller than the first-order shift. The second-order shift is proportional to the square of the atomic velocity and does not depend on the relative directions of the atomic motion and the electromagnetic wave.

Several other factors affect the quality of the information. Atoms in the system may collide with one another; the impacts add noise to the signal. The surrounding environment can perturb the resonance frequencies. Defects in the electronic equipment, stray electromagnetic fields and the ever present thermal radiation all introduce errors. Therefore, a good atomic frequency standard not only must establish a steady, periodic signal but also must minimize these potential errors.



MASTER PENDULUM of this 1920s Shortt clock oscillates in an evacuated enclosure. It actuates an electrical switch to synchronize a slave pendulum, which drives the clock mechanism.

One of the earliest and now widely used methods to sidestep many of these difficulties is called atomic beam resonance, pioneered by I. I. Rabi and his colleagues at Columbia University in the 1930s. The atoms emerge from a small chamber, exit through a narrow aperture and then travel as a beam. The entire instrument can be shielded from stray magnetic and electric fields and insulated from external sources of heat. Perhaps more important, collisions of atoms are virtually eliminated, because the entire device is housed in a long, evacuated chamber. The pressure in the chamber is so low that the atoms are unlikely to strike anything before reaching the other end.

In simplified form, atomic beam resonance involves three steps. The first is to select only those atoms in the appropriate energy level. This selection is accomplished by using a specially shaped magnetic field, which acts as a kind of filter. It allows atoms in one energy level to pass and blocks all others by bending the beam. Only atoms in the correct energy level are bent the correct amount to reach and pass through the aperture that serves as the entrance to the cavity.

The second and crucial step is to send the selected atoms into another energy level. The task is accomplished by passing the atoms through an oscillating microwave field inside a cavity. The atoms will go to another energy level only if the frequency of the applied oscillating microwaves matches their Bohr frequency.

The third step is to detect those atoms that have changed energy levels. At this point, the beam of atoms passes through another magnetic field filter, which allows only atoms in the correct energy level to strike a detector that records the atoms as current flow. An abundance of such atoms will exist if the frequency of the applied oscillating microwaves precisely matches their natural frequency. If the frequency of the applied microwave field is off the mark, fewer atoms change their energy levels, and so fewer will strike the detector. One knows, therefore, that the applied microwaves match the natural frequency of the atoms if the number of atoms striking the detector is maximal. An electronic feedback mechanism, called a servo loop, keeps this value constant. If it finds that the current from the detector is falling off, it changes the frequency of the applied field until the current reaches a maximum again.

By keeping the current from the detector at a maximum, the servo loop maintains the frequency of the applied microwave field at the natural frequency of the atoms. To measure time, one couples the applied field to a frequency divider, which generates timing pulses. By analogy, the atoms represent the quartz crystal in a watch or the master pendulum in a Shortt clock. The applied microwave field is the oscillating circuit or the slave pendulum, which actually drives the clock mechanism.

Minor variations of the atomic beam standard exist. For example, in some devices the atoms that undergo a change in energy level are made to miss, rather than strike, the detector. Not much difference in accuracy exists, however. Rather all the versions to some extent represent trade-offs in terms of size, cost and complexity.

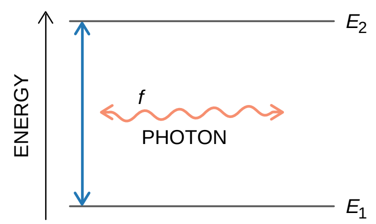
A more important modification of the atomic beam came in 1949, when one of us (Ramsey) invented the so-called separated oscillatory field method. Instead of irradiating the atoms with a single applied field, this technique relies on two fields, separated by some distance along the beam path. Applying the oscillating field in two steps has many benefits, including a narrowing of the resonance and the elimination of the first-order Doppler shift. Jerrold R. Zacharias of the Massachusetts Institute of Technology and Louis Essen and John V. L. Parry of the National Physical Laboratory in Teddington, England, adapted this method to working frequency standards in the mid-1950s.

Currently the separated oscillatory field method provides the most reproducible clocks. The best ones are located at a few national laboratories, although smaller and less accurate versions are commercially available. The clocks rely on cesium, which has several advantages over other elements. It has a relatively high resonance frequency—about 9,192 megahertz—and low resonance width, which lead to an excellent Q. Cesium can also be detected readily and efficiently; all that is needed is a hot metal filament. When a cesium atom strikes the filament, it ionizes and becomes observable as electric current.

The Qs of these standards are about 100 million, exceeding the Q of quartz wristwatches by a factor of several thousand. The greatest reproducibilities are about a part per 10^{14}. The best cesium frequency standards are so much more reproducible than the rate of rotation and revolution of the earth that in 1967 the second was defined as 9,192,631,770 periods of the resonance frequency of the cesium 133 atom.

Resonance Frequency

Atomic frequency standards depend on the quantization of the internal energies of atoms or molecules. A pair of such energy levels, shown here as levels E_1 and E_2, is associated with an atomic resonance. The resonance frequency f, at which it absorbs or emits electromagnetic radiation, is $f = (E_2 - E_1)/h$, where h is Planck's constant. The radiation, however, is not precisely f but instead is spread over a range near f, called Δf. The precision to which f can be measured is proportional to the quality factor, Q, defined by $Q = f/\Delta f$. The higher the Q, the more stable the clock.



One of the most promising improvements in cesium atomic-beam standards is the use of optical pumping to select the atomic states. Beginning in the 1950s optical-pumping techniques were developed by Francis Bitter of M.I.T., Alfred Kastler and Jean Brossel of the École Normale Supérieure and others. In this method, light, rather than a magnetic field, selects atoms in the desired states. Before the atoms are subjected to the microwave field, radiation from a laser is used to drive (or pump) the atoms from one energy level into another. In fact, one can control the number of atoms in energy levels by tuning the frequency of the light.

After the atoms have been irradiated by the microwave field, they pass through a second light beam. Only atoms occupying the correct energy level absorb this light, which they quickly reemit. A light-sensitive detector records the reemissions and converts them into a measurable current. As in atomic beam resonance that relies on magnetic selection, one knows that the applied microwave field matches the natural frequency of the atoms if the current from the detector is at a maximum.

Using light instead of magnets has many advantages. Perhaps the most crucial is that, with the right optical-pumping techniques, all the atoms in the beam can be put into the desired energy level. Magnetic selection merely filters out those that are in the other energy levels. Hence, the signal strength from optical pumping is much higher than it is from magnetic selection. Researchers at various laboratories are developing optically pumped cesium atomic-beam clocks. One such clock, at the National Institute of Standards and Technology (NIST) in Boulder, Colo., has recently become the primary frequency standard for the U.S. Designated NIST-7, it has an expected error of one second in about one million years, making it many times more stable than its predecessor.

There is an optically pumped atomic clock that is available commercially. Such a clock is based on the 6,835-megahertz, hyperfine resonance of rubidium 87. Rather than moving through the apparatus as a beam, the rubidium atoms are contained in a glass cell. The cell also houses a mixture of gases that prevents the rubidium atoms from colliding with the cell walls. A discharge lamp containing rubidium vapor, rather than a laser, irradiates the atoms. A photovoltaic sensor on the opposite side of the cell detects changes in the amount of light absorbed by the atoms. The atoms are prepared, the microwaves applied and the light detected in one cell. As a result, rubidium clocks can be made to fit in a cube about 10 centimeters on a side. In contrast, cesium beam clocks can extend from about 50 centimeters to more than five meters. Rubidium clocks are also much less expensive than are cesium ones.

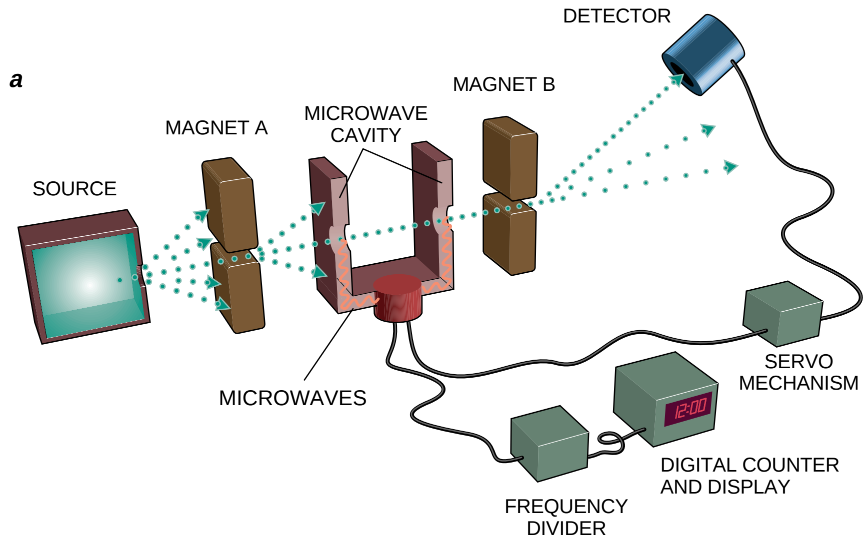
The drawback is that the rubidium devices are generally less accurate and less reproducible. The Q of rubidium standards is about 10 million, a factor of 10 less than the cesium beam's quality factor; their reproducibility is only about a part per 10^{10}. Shifts in the resonance frequency mostly account for the poor reproducibility. The frequent collisions of the rubidium atoms with other gas molecules cause the shifts. But the rubidium standards' short-term stabilities are good—in fact, better than those of some cesium atomic beams.

The atomic clocks described thus far work in a rather roundabout way—by detecting a change in some signal, such as the number of atoms striking a detector, as the frequency of the applied oscillatory field shifts. One way to make use of the radiation emitted by the atoms more directly relies on the principle of the maser (an acronym for microwave amplification by stimulated emission of radiation). In 1953 Charles H. Townes and his associates at Columbia invented the first maser, which was based on ammonia. Beginning in 1960, Ramsey, Daniel Kleppner, now at M.I.T., H. Mark Goldenberg, then at Harvard University, and Robert F. C. Vessot, now at the Harvard-Smithsonian Center for Astrophysics, developed the atomic hydrogen maser, the only type that has been used extensively as an atomic clock.

In this instrument, a radio frequency discharge first splits hydrogen molecules held in a high-pressure bottle into their constituent atoms. The atoms emerge from a small opening in the bottle, forming a beam. Those in the higher energy level are focused by magnetic fields and enter a specially coated storage bulb surrounded by a tuned, resonant cavity.

In the bulb, some of these atoms will drop to a lower energy level, releasing photons of microwave frequency. The photons will stimulate other atoms to fall to a lower energy level, which in turn releases additional microwave photons. In this manner, a self-sustaining microwave field builds up in the bulb—thus the name "maser." The tuned cavity around the bulb helps to redirect photons back into the system to maintain the stimulated emission process. The maser oscillation persists as long as the hydrogen is fed into the system.

A loop of wire in the cavity can detect the oscillation. The microwave field induces a current in the wire, which leads out of the cavity to a series of circuits. The circuits convert the induced current to a lower frequency signal suitable for generating timing pulses.



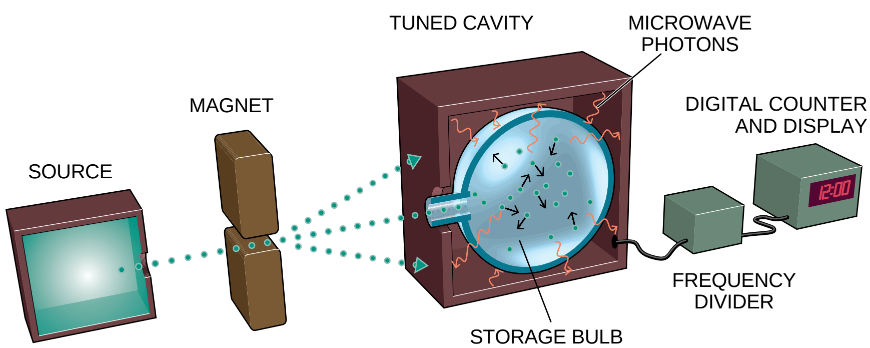
ATOMIC-BEAM frequency standards provide the most accurate, long-term timekeeping. Conventional atomic clocks rely on magnets (*a*). Atoms in the correct energy level are deflected by magnet A through the microwave cavity. Microwave fields oscillating at the resonance frequency of the atoms drive some of them into a second energy level. These atoms are deflected by magnet B so as to strike a detector. The servo mechanism monitors the detector and maintains the frequency of the applied microwaves at the resonance frequency. To keep time, some of the microwaves are

The resonance frequency in the hydrogen maser is about 1,420 megahertz, which is much lower than the resonance frequency of cesium. But because the hydrogen atoms reside in the bulb much longer than cesium atoms do in a beam, the maser's resonance width is much narrower. Consequently, the *Q* of a hydrogen maser standard is about 10^9, exceeding the *Q* of the cesium atomic clock by an order of magnitude. In addition, a hydrogen maser has the highest stability of any frequency standard, better than one part per 10^{15}.

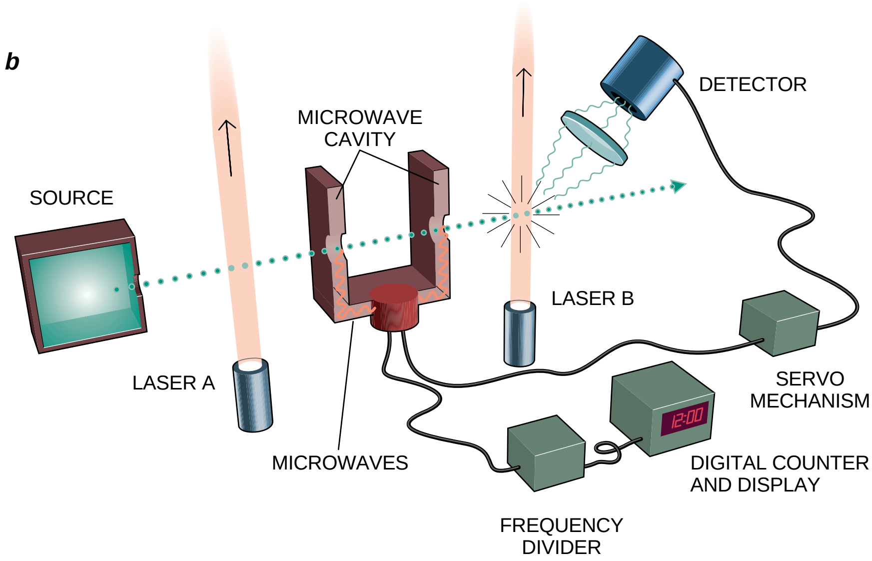
Unfortunately, the maser's superior attributes last just for a few days. Beyond that, its performance falls below that of cesium beams. The stability decreases because of changes in the cavity's resonant frequency. Collisions between the atoms and the bulb shift the frequency by about one part per 10^{11}.

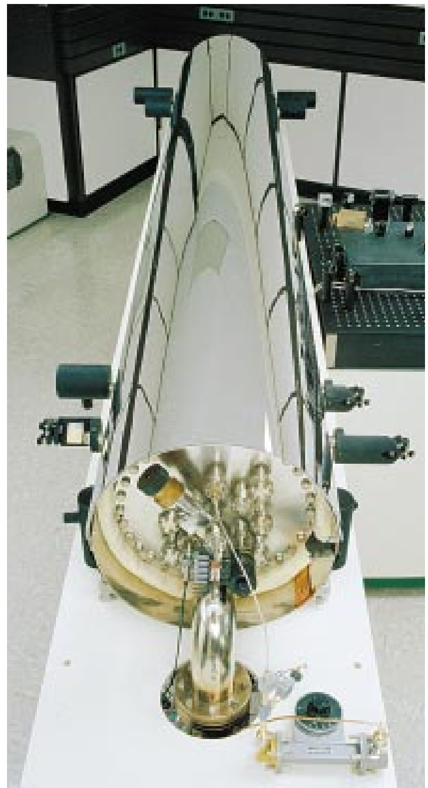
One way to overcome the problem is to operate the hydrogen maser at low temperatures. This condition allows more atoms to be stored (thus resulting in a stronger signal) and reduces electronic noise. Coating the inside of the bulb with superfluid liquid helium also enhances performance. This substance acts as a good surface against which the hydrogen atoms can bounce. More effective magnets, better coating substances and servo loop techniques that keep the cavity resonance centered on the atomic resonance are other approaches now being taken to improve maser stability.

Although the cesium atomic-beam frequency standard is the most accurate, long-term standard we have, several breakthroughs have indicated that it is possible to fabricate even more precise clocks. One of the most promising depends on the resonance frequency of trapped, electrically charged ions. Trapped ions can be suspended in a vacuum so that they are almost perfectly isolated from disturbing influences. The ions themselves stay well separated from one another



ATOMIC HYDROGEN MASER relies on a self-sustaining microwave field to serve as a frequency standard. Hydrogen atoms in the correct energy level are deflected by a magnet into a storage bulb. Some atoms will drop to a lower level, releasing a microwave photon. The photon stimulates other atoms to drop to a lower level, which produces more photons. The process quickly builds up a microwave field in the bulb. The field induces an alternating current in a wire placed in the cavity. The tuned cavity helps to redirect the photons back into the bulb to maintain the process.





directed to a device that divides the frequency into usable timing pulses. Optically pumped standards (*b*) use light rather than magnets to select atoms. Laser A pumps the atoms into the right energy level, preparing them to be excited by the microwaves. Only atoms placed in the correct energy level by the microwaves absorb light from laser B. They quickly reemit that energy, which is sensed by a photodetector. An optically pumped clock using cesium atoms at the National Institute of Standards and Technology, called NIST-7, now keeps time for the U.S. (*photograph*).

because they have the same electric charge. Hence, they do not suffer collisions with other particles or with the walls of the chamber. Ions can be trapped for long periods, sometimes for days.

Two different types of traps are used. In a Penning trap, a combination of static, nonuniform electric fields and a static, uniform magnetic field holds the ions. In a radio frequency trap (often called a Paul trap), an oscillating, nonuniform electric field does the job. Each type of trap has its own characteristic shortcoming. The strong magnetic fields of Penning traps can alter the resonance frequency. The electric field in Paul traps can create heating effects that cause Doppler shifts. The kind of trap chosen depends on its suitability for a particular experimental setup.

Workers at Hewlett-Packard, the Jet Propulsion Laboratory in Pasadena, Calif., and elsewhere have fabricated experimental standard devices using Paul traps. The particles trapped were mercury 199 ions. This ion was selected because it has the highest hyperfine frequency—40.5 gigahertz—of all the atoms that are appropriate for the trapping technique. A few million such ions are caught between the electric fields generated by electrodes. Then the ions are optically pumped by ultraviolet radiation from a lamp. Subsequent operation resembles that of the optically pumped standards, but the maximum Qs of trapped-ion standards exceed 10^{12}. This value is 10,000 times greater than that for current cesium beam clocks. Their short-term stabilities are also extremely good, although they do not yet reach those of hydrogen masers. The second-order Doppler shift limits the reproducibility to about one part per 10^{13}.

The Doppler shifts can be greatly reduced by laser cooling. In 1975 David J. Wineland, now at NIST, Hans G. Dehmelt of the University of Washington, Theodor W. Hänsch, now at the University of Munich, and Arthur L. Schawlow of Stanford University first proposed such a technique. In essence, a beam of laser light is used to reduce the velocities of the ions. Particles directed against the laser beam absorb some of the laser photon's momentum. As a result, the particles slow down. To compensate for the Doppler shifting as the particle moves against the laser, one tunes the beam to a frequency slightly lower than that produced by a strongly allowed resonance transition.

Many laboratories are developing frequency standards based on laser-cooled ions in traps. A standard based on beryllium 9 ions, laser-cooled in a Penning trap, has been constructed. Its reproducibility is about one part per 10^{13}, limited as it is by collisions of the ions with neutral molecules. Improvements in the quality of the vacuum should significantly increase the reproducibility because the uncertainty of the second-order Doppler shift is only about five parts per 10^{15}.

During the past few years, there have been spectacular developments in trapping and cooling neutral atoms, which had been more difficult to achieve than trapping ions. Particularly effective laser cooling results from the use of three pairs of oppositely directed laser-cooling beams along three mutually perpendicular paths. A moving atom is then slowed down in whatever direction it moves. This effect gives rise to the designation "optical molasses." Several investigators have contributed to this breakthrough, including William D. Phillips of NIST in Gaithersburg, Md., Claude Cohen-Tannoudji and Jean Dalibard of the École Normale Supérieure and Steven Chu of Stanford [see "Laser Trapping of Neutral Particles," by Steven Chu; SCIENTIFIC AMERICAN, February 1992].

Neutral-atom traps can store higher densities of atoms than can ion traps, because ions, being electrically charged, are kept apart by their mutual repulsion. Other things being equal, a larger number of atoms results in a higher signal-to-noise ratio.

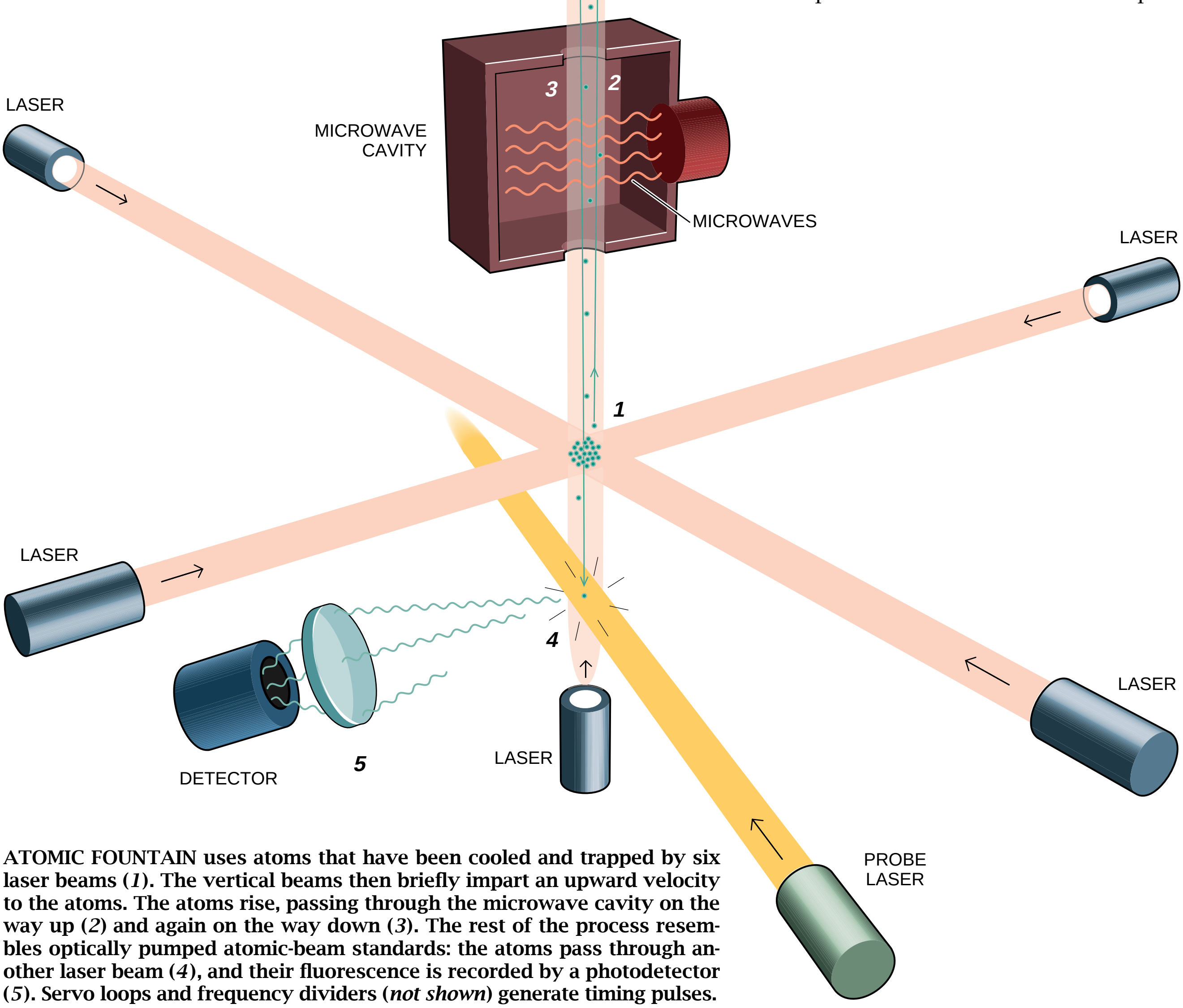
The main hurdle in using neutral atoms as frequency standards is that the resonances of atoms in a trap are strongly affected by the laser fields. A device called the atomic fountain surmounts the difficulty. The traps capture and cool a sample of atoms that are then given a lift upward so that they move into a region free of laser light. The atoms then fall back down under the influence of gravity. On the way up and again on the way down, the atoms pass through an oscillatory field. In this way, resonance transitions are induced, just as they are in the separated oscillatory field beam apparatus.

The *Q* of such a device can be higher than that of an atomic beam because the time between the two passages can be longer. Experiments on atomic fountains have been done by Chu and his co-workers at Stanford and by André Clairon of the Primary Laboratory of Time and Frequency in Paris and Christophe Salomon of the École Normale Supérieure and their co-workers.

Much current research is directed toward laser-cooled ions in traps that resonate in the optical realm, where frequencies are many thousands of gigahertz. Such standards provide a promising basis for accurate clocks because of their high *Q*. Investigators at NIST have observed a *Q* of 10^{13} in the ultraviolet resonance of a single laser-cooled, trapped ion. This value is the highest *Q* that has ever been seen in an optical or microwave atomic resonance. Because of technical difficulties, none of the optical frequency clocks constructed so far, however, has operated over extended periods.

The variety of high-performance frequency standards that exist today might seem to obviate the need for future devices of even greater performance. After all, current atomic clocks are so accurate that they have redefined some of our basic units. As mentioned earlier, the second is now based on the resonance frequency of the cesium atom. Also by international agreement, the meter is defined as the distance light travels in 1/299,792,458 of a second. The voltage unit is maintained by the characteristic frequency associated with a voltage that appears in a so-called Josephson junction in a superconducting circuit.

There are, however, applications that tax the capacity of modern clocks. Radio astronomy is a good example. Astronomers often use several telescopes spaced thousands of kilometers apart



ATOMIC FOUNTAIN uses atoms that have been cooled and trapped by six laser beams (*1*). The vertical beams then briefly impart an upward velocity to the atoms. The atoms rise, passing through the microwave cavity on the way up (*2*) and again on the way down (*3*). The rest of the process resembles optically pumped atomic-beam standards: the atoms pass through another laser beam (*4*), and their fluorescence is recorded by a photodetector (*5*). Servo loops and frequency dividers (*not shown*) generate timing pulses.

to study a stellar object, a technique that dramatically increases the resolution [see "Radio Astronomy by Very-Long-Baseline Interferometry," by Anthony C. S. Redhead; SCIENTIFIC AMERICAN, June 1982]. Two radio telescopes spaced 10,000 kilometers apart have an effective angular resolution more than one million times better than either telescope alone. But to combine the data from each telescope appropriately, investigators need to know precisely when each telescope received the signal. Present-day hydrogen masers have the stability required for such observations. More stable clocks may be needed for space-borne radio telescopes.

Highly stable clocks are essential for the best tests of relativity. Timing measurements of millisecond pulsars, some of which are as stable as the best atomic clocks, offer evidence for gravity waves. In 1978 Joseph H. Taylor, Jr., and his associates at Princeton University found that the period of a binary-pulsar system has been slowly varying by just the amount that would be expected for the loss of energy by gravitational radiation, as predicted by general relativity. Greater precision can be achieved if measurements are taken over many years, so clocks with better long-term stability would be useful.

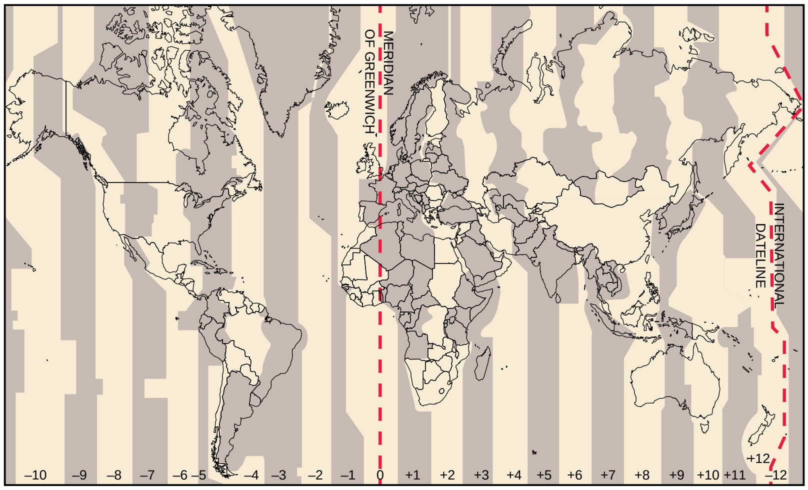
In other tests of relativity, Vessot and his colleagues confirmed the predicted increase in clock rates at high altitudes. They sent on board a rocket a hydrogen maser and measured the small, relativistic clock shift to within an accuracy of 0.007 percent at an altitude of 10,000 kilometers. Highly stable clocks have also been used by Irwin I. Shapiro, now at the Harvard-Smithsonian Center for Astrophysics, to observe the relativistic delay of a light signal passing by the sun.

Ultraprecise timekeeping has more practical applications as well—most notably, for navigation. The location of *Voyager 2* as it sped by Neptune was determined by its distance from each of three widely separated radar telescopes. Each of these distances in turn was obtained from accurate measurements of the eight hours it took for light to travel from each telescope to the spacecraft and return.

Navigation is, of course, also important on the earth. One of the latest applications of precise clocks is the satellite-based assemblage called the Global Positioning System, or GPS. This system relies on atomic clocks on board orbiting satellites. The GPS enables anyone with a suitable radio receiver and computer to determine his or her position to approximately 10 meters and the correct time to better than 10^{-7} second. Two observers monitoring the same satellite can synchronize their clocks to within a few nanoseconds.

It is expected that the GPS will have widespread practical applications, such as pinpointing the positions of ships, airplanes and even private automobiles. The GPS was used during the 1991 Persian Gulf War to enable troops to determine their positions on the desert. Commercial receivers can be purchased for less than $1,000, although these civilian versions are limited to an accuracy of about 100 meters because of deliberate scrambling of the signals transmitted from the satellites. A full complement of 24 satellites would give 24-hour, worldwide coverage. The system is nearly complete.

These and other applications show the importance of time and frequency standards. The anticipated improvements in standards will increase the effectiveness of the current uses and open the way for new functions. Only time will tell what these uses will be.



Coordinating Time Scales

In the article, we discuss the measurement of an interval of time, such as a second or a minute. This process requires only a good clock. But to be able to state that an event happened at a particular time, say, 22 seconds after 12:31 P.M. on July 5, 1993, requires synchronization with a clock that is, by mutual agreement, the standard. The world's "standard clock" exists on paper as an average of the best clocks in the world. The International Bureau of Weights and Measures in Sèvres, France, is responsible for coordinating international time. This coordinated time scale is called International Atomic Time, or TAI.

Many users require a time scale that keeps pace with the rotation of the earth. That is, averaged over a year, the sun should be at its zenith in Greenwich, England, at noon. The day as determined by the apparent position of the sun is irregular but on the average longer than the 24 hours as defined by TAI. To compensate, another time scale, called Coordinated Universal Time, or UTC, is specified by occasionally adding or subtracting a whole number of seconds from TAI. These seconds, or leap seconds, are inserted or deleted, usually on December 31 or June 30, to keep UTC within 0.9 second of the time as defined by the rotation of the earth. The record of leap seconds must be consulted to determine the exact interval between two stated times.

FURTHER READING

FROM SUNDIALS TO ATOMIC CLOCKS: UNDERSTANDING TIME AND FREQUENCY. J. Jespersen and J. Fitz-Randolph. Dover, 1982.

HISTORY OF ATOMIC CLOCKS. N. F. Ramsey in *Journal of Research of the National Bureau of Standards,* Vol. 88, No. 5, pages 301–320; September/October 1983.

PRECISE MEASUREMENT OF TIME. N. F. Ramsey in *American Scientist,* Vol. 76, No. 1, pages 42–49; January/February 1988.

TIME AND FREQUENCY. Special issue of *Proceedings of the IEEE,* Vol. 79, No. 7; July 1991.

Surgical Treatment of Cardiac Arrhythmias

To save the life of a doomed patient, the author and his colleagues developed a now standard surgical procedure for correcting lethally fast heartbeats in many people susceptible to them

by Alden H. Harken

In 1978 a vice president of a bank in Philadelphia collapsed at work when his heart began to beat dangerously fast. Fortunately, his co-workers were able to administer cardiopulmonary resuscitation immediately, keeping him alive until emergency medical workers arrived. He was soon brought to the Hospital of the University of Pennsylvania, where I was a junior member of the surgical faculty.

Little did either of us know that within weeks of this episode we would participate together in making a small piece of surgical history. Desperate to prevent the banker's imminent death, my colleagues and I devised a new surgical treatment to correct the underlying disturbance that caused his heart to malfunction. Since then, hundreds of other patients have been aided by this therapy. At the same time, further research has expanded insight into why our treatment strategy, born of necessity, proved so useful.

I well remember our initial evaluation of the banker's medical condition because we were in for a surprise. When he first appeared at the hospital, we suspected he had suffered a heart attack (myocardial infarction): the death of cardiac muscle after blockage of an artery feeding that tissue. But tests told a different story. Indeed, the muscle was in good shape, except for a small area that had been damaged during a heart attack several years before.

His heart had malfunctioned now because it became suddenly and lethally unstable electrically. The electrical wiring system that regulates the heartbeat induces the cardiac muscle to contract and thus push blood into the arterial circulation some 72 times a minute. The man's muscle had begun to receive much more frequent signals, leading to abnormally fast pumping. If the heart beats too rapidly, its interior chambers do not have time to fill with blood. Because the organ cannot eject something it does not receive, delivery of blood to the body's tissues, including to the cardiac muscle itself, can drop precipitously, causing the heart to stop.

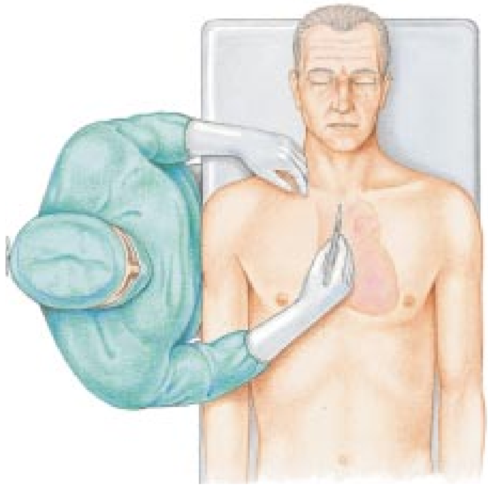
Although we had originally expected to find evidence of a new heart attack, we were also aware that the banker's electrical derangement was not unique. Six years earlier Hein J. J. Wellens, then at the University of Limburg in the Netherlands, observed that excessively fast pumping occurred in certain patients months or years after a heart attack.

We understood as well that medications designed to prevent arrhythmias, or abnormal heartbeats, could restore proper functioning in some people, and so we tried every type available. Each failed. In a span of three weeks at the hospital, the banker seriously extended his metaphysical line of credit, suffering three additional cardiac arrests. To let him leave under those conditions would most assuredly have been fatal—and he knew it.

At the time, I was privileged to be working with Mark E. Josephson and Leonard N. Horowitz, who specialized in diagnosing cardiac electrical abnormalities. They concluded that the banker's trouble stemmed from a disturbance known as a reentrant electrical circuit in the heart. That being the case, we thought we might be able to interrupt the circuit surgically.

To follow our logic, it helps to know a bit about how the heart's electrical system controls cardiac activity. The heart, which is divided into four chambers, is essentially a ball of muscle (myocardium) lined by conduction tissue: unique fibers that form a kind of internal nervous system. These special fibers convey electrical impuls-

ALDEN H. HARKEN, a practicing cardiac surgeon, is professor and chairman of the department of surgery at the University of Colorado Health Sciences Center in Denver. He earned his M.D. from Case Western Reserve School of Medicine in 1967. After completing his surgical residency at Peter Bent Brigham Hospital and Children's Hospital, both in Boston, he joined the Hospital of the University of Pennsylvania in 1976. Harken has held his current posts since 1984.



LIFESAVING OPERATION involves excising flap of diseased muscle (*lined area in image at right*), about three square centimeters in area and several millimeters thick, from the inner surface of a patient's heart. When successful, the surgery halts propagation of impulses through a pathway known as a reentrant circuit, which may arise months or years after a heart attack and can fatally disturb normal cardiac rhythms. The surgeon has entered the left ventricle through an incision (*broken line in inset*) in dead scar tissue (*shaded area in inset*) left by the heart attack. Clamps hold back the edges of the incision.

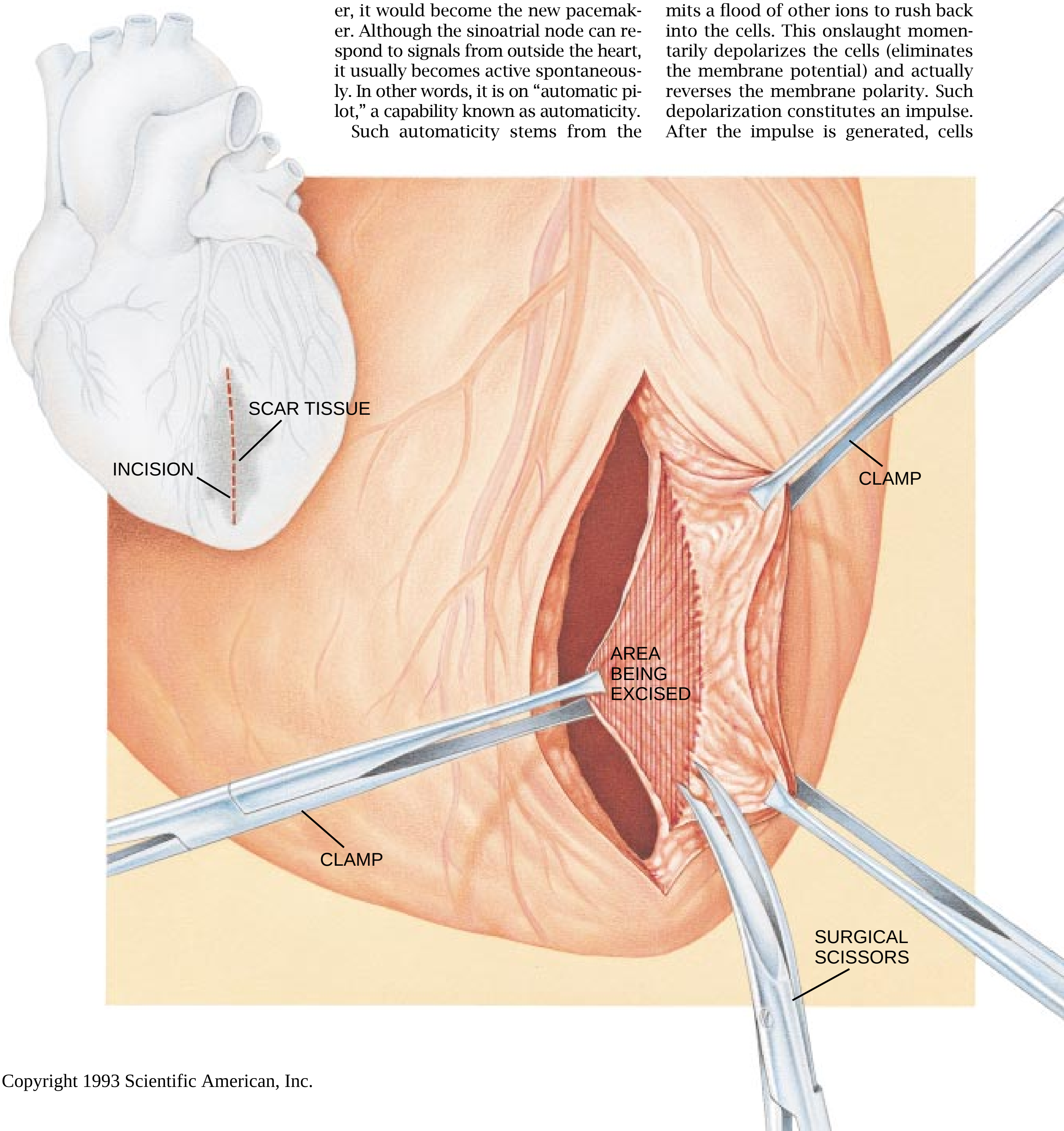
es swiftly to the entire cardiac muscle.

In response to the impulses, the muscle contracts—first at the top of the heart and slightly thereafter at the bottom. As contraction begins, oxygen-depleted, venous blood is squeezed out of the right atrium (one of two small upper chambers) and into the larger right ventricle below. Then the ventricle ejects the blood into the pulmonary circulation, which resupplies oxygen and delivers the blood to the left side of the heart. In parallel with the events on the right, the muscle pumps newly oxygenated blood from the left atrium into the left ventricle and, from there, out to the aorta, which distributes it to every part of the body.

The signal giving rise to these machinations emanates from a cluster of conduction tissue cells collectively known as the sinoatrial node. This node, located at the top of the right atrium, establishes the tempo of the heartbeat; hence, it is often referred to as the cardiac pacemaker. It sets the tempo simply because it issues impulses more frequently than do other cardiac regions, once about every 830 milliseconds. If something provoked another part of the heart to fire at a faster rate, as occurred in the banker, it would become the new pacemaker. Although the sinoatrial node can respond to signals from outside the heart, it usually becomes active spontaneously. In other words, it is on "automatic pilot," a capability known as automaticity.

Such automaticity stems from the unique leakiness of the membrane encasing nodal cells. As is true of the membrane surrounding muscle cells and neurons, the nodal cell membrane is studded with pumps that transport ions into and out of the cell. The net result of this exchange is the creation of an electrical potential, or unequal charge distribution, across the membrane. Yet unlike muscle and nerve cells, which maintain their resting potential until they are jogged by an outside stimulus, nodal cells allow certain ions to leak back out of the cells. This outflow reduces the membrane potential to a critical value.

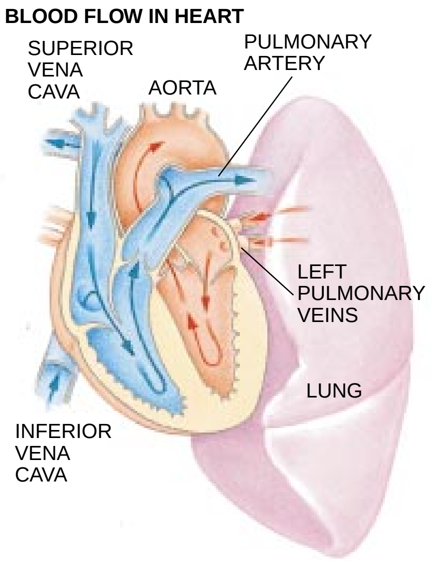
At that point, the membrane permits a flood of other ions to rush back into the cells. This onslaught momentarily depolarizes the cells (eliminates the membrane potential) and actually reverses the membrane polarity. Such depolarization constitutes an impulse. After the impulse is generated, cells

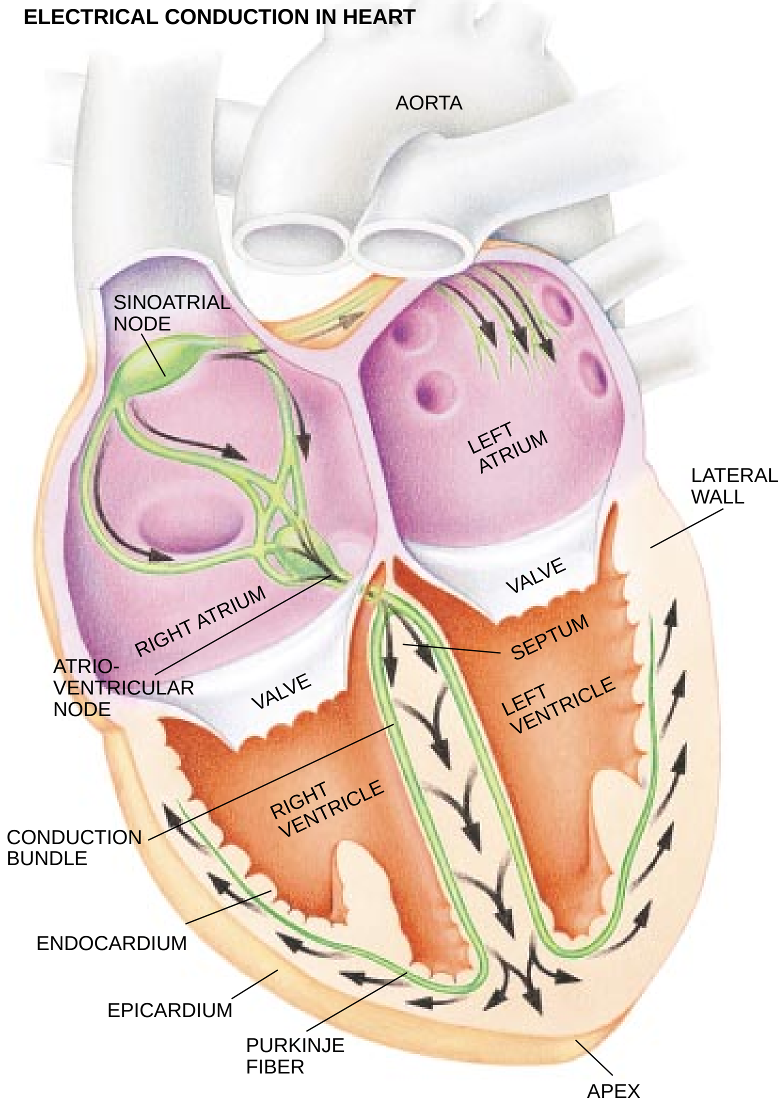


repolarize and prepare for firing anew.

Impulses born at a cell in the sinoatrial node typically speed instantly through the rest of the node; from there, they course through the entire heart in the span of 160 to 200 milliseconds. Traveling along conduction fibers, they first race across both atria and then regroup at the atrioventricular node, a cellular cluster centrally located atop the ventricles. After a pause, they course down the ventricles along a conduction cable that divides into two branches known as conduction bundles; these further ramify to form arbors of thinner projections called Purkinje fibers. One arborized bundle serves each ventricle, sending signals first along the surface of the septum (a wall dividing the two ventricles) to the tip of the heart (the apex) and, from there, up along the inner surface of the external (lateral) walls to the top of the ventricle.

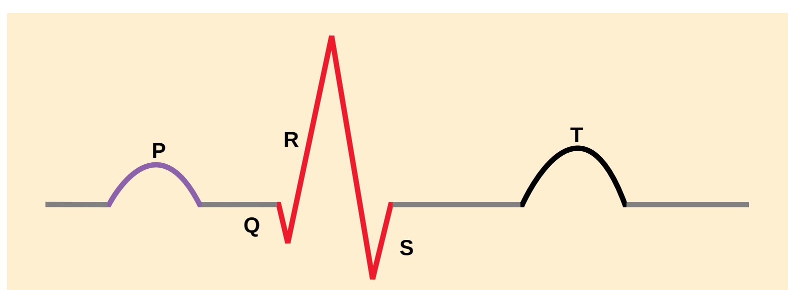
As impulses from the conduction fibers reach muscle, they activate the overlying cells. Muscle cells, too, are capable of relaying impulses, albeit more slowly than do conduction fibers. The cells of the endocardium (the inner surface of the wall) depolarize first and relay





The Making of a Heartbeat

A specialized electrical conduction system (*green in large heart*) normally regulates the steady beating of the heart. The impulses (*black arrows in image at right*) that induce pumping are issued at set intervals from the sinoatrial node (*large green oval at top left*), or the cardiac "pacemaker." From there, they race to the atrioventricular node (*above the ventricles*) and, after a brief pause, speed down along the septum to the bottom of the heart and up its sides. Meanwhile the impulses also migrate from the conduction fibers across the overlying muscle, from the endocardium to the epicardium, thereby triggering the contractions that force blood (*arrows in small diagram above*) through the heart and into the arterial circulation. The spread of electricity through a healthy heart gives rise to the familiar electrocardiogram at the bottom right. The P wave (*purple*) and QRS wave (*red*) form as impulses pass through the atria and ventricles, respectively; the T wave (*black*) arises as cardiac cells, which cannot be stimulated for a while after they fire, recover their excitability.



the impulses through the thickness of the muscle to the outer surface (the epicardium). Depolarization, in turn, triggers contraction.

Josephson and Horowitz suggested that diseased cells had distorted this normal flow of electricity in the banker's heart. After a heart attack, many cells surrounding the resulting scar (the group of cells killed by lack of blood delivery) continue to live but are abnormal electrically; they may conduct impulses unusually slowly or fire when they would typically be silent.

These diseased areas, my co-workers indicated, might perturb smooth signaling by forming a reentrant circuit in the muscle: a pathway of electrical conduction through which impulses can cycle repeatedly without dying out. In our patient's case, the circuit was thought to be in the left ventricle, where his heart attack, in common with most others, occurred. (Activation of reentrant circuits some time after a heart attack is now believed to take place in a sizable number, perhaps 10 percent, of the roughly 1.2 million Americans who suffer heart attacks every year.)

Passage of impulses through a reentrant circuit can be envisioned by imagining a wave of impulses encountering, say, the bottom of an oval scar in the left ventricle. On reaching the scar, the wave would split in two, to detour around both sides of the dead area. If diseased cells somehow interrupted impulses propagating along one of those branches, impulses might still flow up along the opposite branch and over the top of the oval. Then they might traverse the previously blocked path and return to the beginning of the circuit—a region we call the origin.

If this circuit were negotiated slowly enough, the origin would have repolarized and become responsive once again to stimulation. (Between the time cells depolarize and repolarize, they are generally refractory, or incapable of responding to new impulses.) In that case, the impulses could reexcite the origin, sending impulses back into the diseased circuit and also out to the rest of the ventricular muscle. Despite the slow conduction, the impulses could complete the circuit in a shorter time than the interval between normal heartbeats. Hence, persistent cycling could enable the origin of the circuit to become the new pacemaker and to provoke sustained ventricular tachycardia: excessively rapid pumping by the ventricles.

We knew that continuous passage through reentrant circuits could occur in humans because Wellens had established that fact in the 1970s. Fortunately for us, he also introduced a procedure for determining whether a quiescent circuit lurks in a patient who survives a life-threatening episode of tachycardia and whether any existing drugs can prevent renewed activation of the pathway. A physician threads an electrode known as a pacing catheter into the heart and issues a series of specifically timed impulses. Initiation of sustained premature heartbeats confirms that a patient harbors a reentrant pathway. (In contrast, impulses delivered to a healthy heart would yield only single contractions that would not be repeated.) Next, the individual is given an antiarrhythmic drug. If paced stimuli now fail to trigger sustained tachycardia, the finding implies the drug should be helpful.

When Josephson and Horowitz performed the procedure on the banker, they found they could indeed induce persistent tachycardia and that, sadly, no antiarrhythmic medications could aid him. I met with the two of them soon afterward in their tiny, windowless catheterization laboratory. Knowing our patient carried a life-threatening electrical pathway inside his heart, we began wondering if we might prevent its activation by surgically removing all or part of the culprit circuit, especially the origin. We realized the plan could fail, or that by removing the tissue, we might actually create other problems. But we were out of options.

Before proceeding, we had to develop a way to locate the renegade pacemaker. We hoped we might find it by analyzing signals reaching an electrode placed directly on the inner or outer surface of the heart [*see bottom illustration on next page*]. More specifically, we planned to induce sustained tachycardia with a pacing electrode. During each heartbeat, we would measure electric currents produced at a single site (consisting of a small cluster of cells) along the diseased border of the heart attack scar. We would start at a position arbitrarily designated as 12 o'clock and proceed around the "clock face" back to the beginning.

We would delineate the circuit by comparing the time of electrical activation in each region against that seen in healthy tissue. Regions that generated currents before the healthy tissue did would be revealed as belonging to the circuit; the area that became excited earliest would be the pacemaker. We could not rely on standard electrocardiography for this purpose because it lacked the specificity we needed. Familiar electrocardiogram tracings, made by attaching electrodes to the skin, reflect the summed activity of many thousands of cells in the heart; they cannot identify the precise swatch of muscle that is depolarized at any given moment.

Our general approach made sense, but no one had ever attempted to "map" the flow of signals in the living, pumping chambers of the human heart by recording directly from the organ's surface. We had no idea whether we could obtain decipherable results. The next day I was scheduled to remove a cancerous lung from a different patient. He kindly agreed to let us try to detect signals directly from the outside of his heart. To our delight, we could clearly discern when a wave of impulses crossed any point on the muscle.

I was now ready to discuss our proposed strategy with the banker. Not knowing whether the origin of the circuit—the zone of earliest activation—was closer to the inside or outside of the cardiac muscle, we intended to map both the inner and outer surfaces. We planned to reach the interior by opening the heart through the existing scar. (Cutting into healthy tissue would, after all, destroy new tissue unnecessarily.) If we found the troublesome region, we proposed to remove it surgically. To keep blood moving through the patient's body during the operation, we should have to attach him to a heart-lung machine. This device diverts unoxygenated blood into an artificial lung. Blood containing oxygen is then pumped back into the arterial circulation via the aorta.

People often call physicians "courageous," but it was our patient who was brave. After I described our therapeutic strategy in great detail, he posed the dreaded question: "How many times have you done this before?" I told him, "Never." Then he asked how many times anyone had performed the operation previously. I informed him it was untried. Despite these unsettling answers, he gave me a confident smile and said, "Go ahead."

The next morning we were able to pinpoint and excise the region of earliest activity, which turned out to reside on the inside surface. (Today we know that virtually all reentrant pathways weave through cells in or close to the endocardium.) Our patient not only resumed banking but also went on to become the county tax assessor. I lost track of him a few years ago, but as of a decade after our treatment, he had suffered no further arrhythmias.

Not everyone who has the surgery is as lucky as the banker was, however. Of all the patients who undergo the procedure after surviving an episode of persistent tachycardia, approximately 9 percent succumb either during the operation or within a month after it. On the other hand, 80 percent of surgically treated patients live for at least a year without recurrence of tachycardia, and 60 percent survive for five years or more. The candidates most likely to do well are those whose heart muscle is damaged least.

In addition to assembling survival statistics, we have discovered since 1978 that reentrant pathways need not be as large as we originally thought. Those occurring at a microscopic level can be equally pernicious. In fact, microanatomic reentrant circuits seem to be the most common form of all.

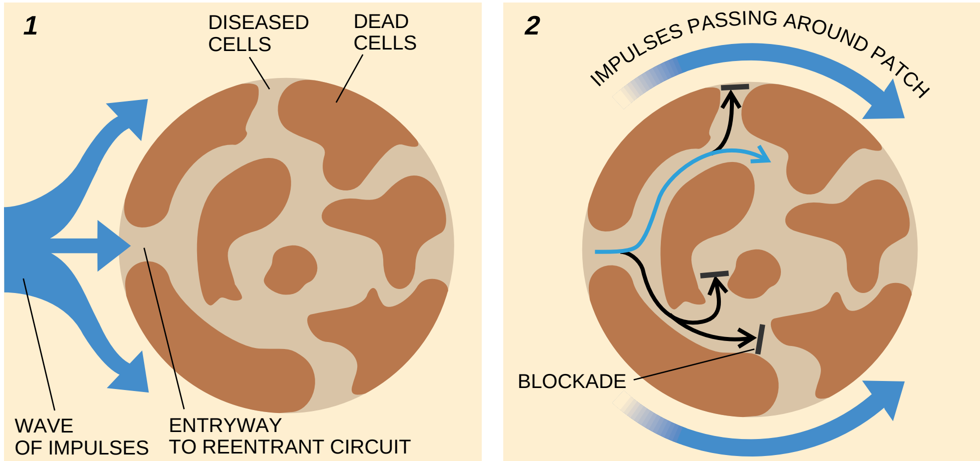
The notion that microcircuits could exist was first suggested in the early 1970s by another surgeon: James L. Cox, then at Duke University. He argued that a small bit of mottled tissue, consisting of diseased cells interspersed with islands of dead cells, could set up the conditions needed to establish reentrant tachycardia. In such a microscopic circuit, impulses that encounter a divided pathway at an entryway to a mottled patch would split and travel along both routes [*see top illustration on this page*]. As is true of larger, "macro" reentrant circuits, impulses propagating along one branch would encounter a one-way blockade. At the same time, impulses flowing along the other branch would meander through a maze of diseased cells and return along the previously blocked lane.

If conduction through the diseased tissue were sufficiently slow, the impulses would come back to the entryway, or origin of the circuit, after that site was no longer refractory. Excitation of the site would then stimulate the ventricular muscle to contract and, at the same time, would send the impulses back into the microcircuit again and again. Instead of traveling along the circumference of a scar, then, a reentrant circuit could trace a recursive path through a more localized maze of cells in the diseased boundary between a heart attack scar and fully healthy tissue.

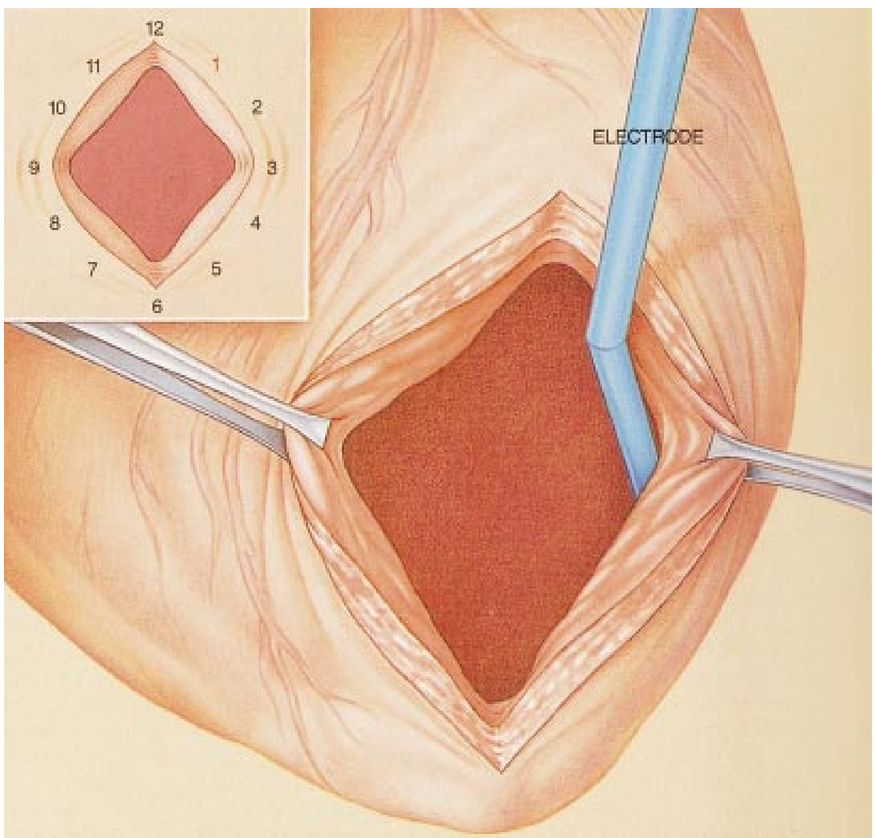
Two of my colleagues, Glenn J. R. Whitman and Michael A. Grosso, decided to test this idea in the early 1980s. They were able to create small heterogeneous zones consisting of mixed dead and living but diseased cells in the ventricles of test animals. These animals, not previously susceptible to the electrical induction of self-sustaining tachycardia, became highly prone to it.

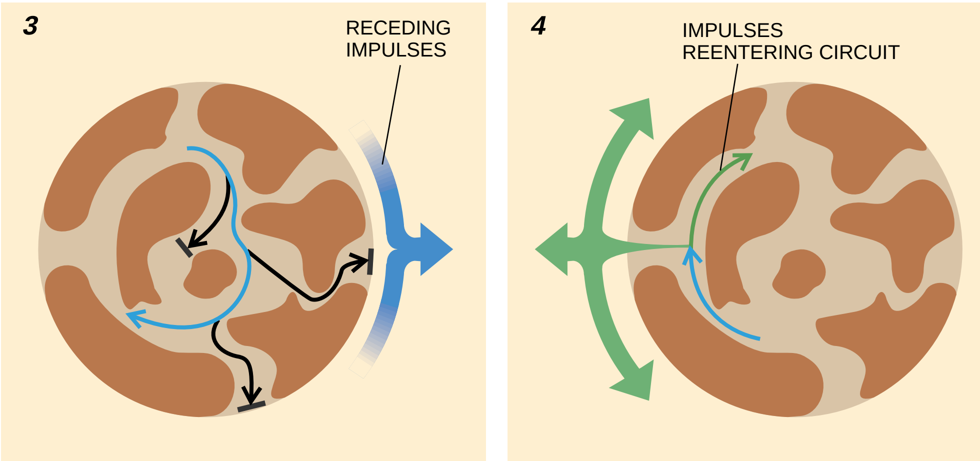
Whitman and Grosso assumed that if the mottled tissue were at fault, killing all the cells in the patch should restore appropriate electrical activity in the heart. Instead of wandering through a dangerous maze, impulses encountering the homogeneous patch of killed tissue would either be extinguished or zoom around it through adjacent healthy cells. Sure enough, when the mottled patches were destroyed, the predisposition to arrhythmia vanished.

These findings revealed that mottling could set the stage for reentrant tachycardia. They also provided the hindsight needed to explain why a different surgical treatment tested by us and others in various patients had not worked well. Believing that the scar itself was somehow responsible for the electrical disturbances, we had previously removed only the dead tissue. Whitman and Grosso's work indicated that this approach was doomed to failure because it left the true culprit—the zone of mixed living and dead cells—in place.



REENTRANT CIRCUIT, a closed loop through which impulses can cycle repeatedly, has formed in a patch of cardiac muscle populated by diseased tissue (*beige*) and islands of dead cells (*brown*). Continuous electrical activity in such a circuit can usurp the normal pacemaker function of the sinoatrial node, leading to ventricular tachycardia (a dangerously fast heartbeat). Persistent cycling begins when a wave of impulses (*blue arrows in 1*) enters the patch and divides (*2*) at the entryway. If





signals flowing along one pathway (*two-headed black arrow*) encounter a temporary blockade (*vertical lines*), those propagating along a second pathway (*thin blue arrow*) may return to the entryway (*3*), or origin, through the previously blocked alley. If the entryway has regained excitability (*4*), the returning impulses will re-excite that area. They may thus fan out through healthy muscle (*large green arrows*) and proceed back through the worrisome circuit again and again (*thin green arrow*).

Yet we still faced two significant puzzles, one scientific and one clinical. Why is it that reentrant circuits do not become active every time the heart beats in susceptible patients? In other words, why can people often survive for months or years before deadly disturbances of rhythm arise? We also wondered how we might noninvasively identify patients at risk for reentrant tachycardia before they experienced a potentially life-threatening episode.

The simplistic explanation for why a reentrant circuit does not jump into action with each heartbeat seemed to be that impulses fired by the sinoatrial node cannot cycle repeatedly through the troublesome pathway. At the end of the first cycle, they return to a still refractory starting site. Blocked from reentering the circuit, they go no further. Unfortunately, this explanation did not clarify how persistent cycling does arise. We now think it is triggered when, in a case of exquisite bad luck, an electrically irritable cell lying adjacent to a reentrant pathway fires spontaneously in a narrow window of time between one activation of the sinoatrial and atrioventricular nodes and the next.

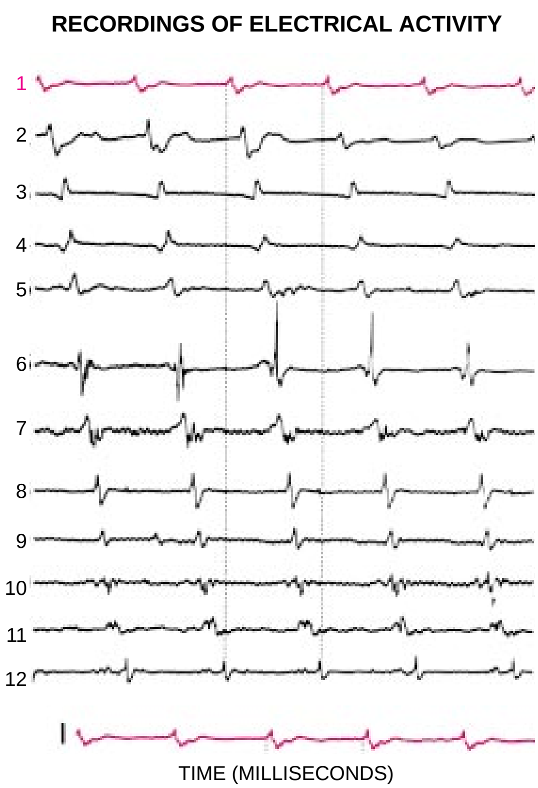
We came to this conclusion after reviewing research reported in the late 1970s by our colleagues E. Neil Moore and Joseph F. Spear of the Hospital of the University of Pennsylvania. By impaling cells on tiny, needlelike electrodes, Moore and Spear were able to track changes in the membrane potentials of single, diseased cardiac cells taken from the area surrounding heart attack scars. After healthy cells depolarize, they repolarize smoothly. In the diseased cells, by contrast, the membrane potential fluctuated markedly during the repolarization period.

We presumed that these fluctuations would sometimes progress to premature depolarization, or firing of an impulse. If an irritable cell happened to lie next to a reentrant pathway, it might well insert an impulse into the worrisome channel during the interval between normal heartbeats.

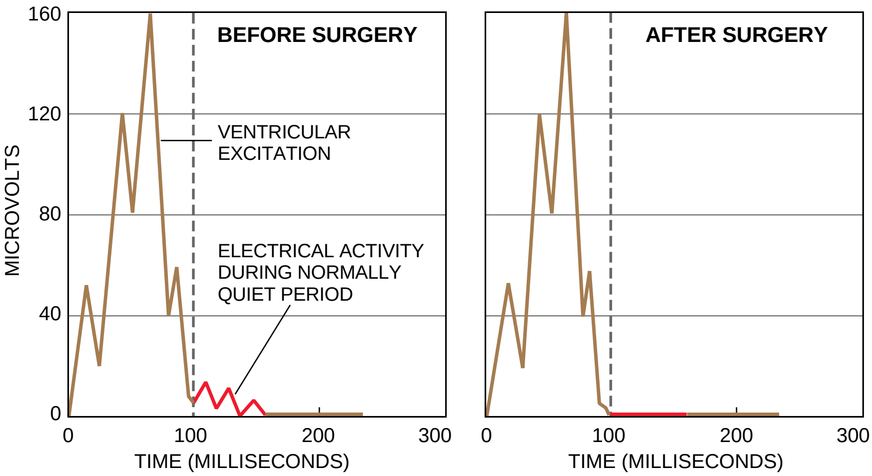
This insertion might activate a reentrant circuit, whereas an impulse originating at the sinoatrial node would not, because recent passage of an impulse through a pathway can alter the electrochemical characteristics of that pathway and slow conduction of a subsequent signal. Thus, the impulse delivered by the irritable cell could pass through the circuit more slowly than would a prior signal originating at the sinoatrial node. If delivery of the wayward impulse were timed properly, the impulse propagating through the circuit would return to the entryway at a most devastating moment: after the site regained excitability (and so could relay the impulse onward) but before the sinoatrial node fired for a second time (thereby gaining control of the heartbeat). Hitting a receptive target, the impulse might proceed to run many unimpeded laps around the lethal circuit.

Our second problem—readily identifying patients at risk for reentrant tachycardia—was resolved masterfully by our co-worker Michael B. Simson, a person of many talents. Aside from being a superb cardiologist, he is, as I sometimes say, an enthusiastic sports-car hack and computer driver. Steering his beat-up sports car home one night after sitting in on one of our surgical research meetings, he began to ponder the electrical noise, or seemingly random signals, emanating from the hood of his car. If he simply monitored the currents reaching the hood, he reasoned, the resulting data would be indecipherably chaotic. But if he wanted to track the electrical impulses coming specifically from his distributor, he might well discern them by signal averaging.

In this procedure, he would record the voltage and direction (the electrical vector) of currents flowing toward and



CARDIAC MAPPING—performed by moving an electrode (*blue*) over diseased tissue surrounding a heart attack scar—enables a surgeon to identify a reentrant circuit. In one patient, recordings of electrical activity (*tracings at right*) made at 12 sites (*inset at left*) around the inner surface during induced ventricular tachycardia delineated a large, recurring electrical circuit that began at the one o'clock position (*tracing 1*) and progressed clockwise around the circumference of the scar. A surgeon can block impulse propagation in such a circuit by excising the region of earliest excitation, which can be assumed to be the entryway to the circuit.



SPECIAL ELECTROCARDIOGRAMS made by a technique called signal averaging were produced before (*left*) and after (*right*) a patient underwent surgery to prevent the continuous passage of impulses through a reentrant pathway. The jagged red line at the left reflects slow, self-limited electrical conduction through the worrisome pathway during the normally quiet period immediately following ventricular excitation. Flattening of the line after surgery means slow conduction through the potentially deadly circuit has been eliminated. Signal-averaged electrocardiograms can now identify patients at risk for reentrant arrhythmias.

away from the hood during particular phases of rotation by his distributor rotor. If he summed the signals obtained by repeated measurements in a given phase, random currents would tend to cancel one another out, leaving a record of only those produced by the rotor. Dividing the result by the number of readings made in a selected phase would give him a measure of the current generated by the distributor in that phase.

It then occurred to Simson that he might apply much the same approach to screen heart attack victims for susceptibility to reentrant tachycardia. Perhaps signal averaging would enable him to detect very slow electrical activity persisting after the normal flow of signals passed through the ventricles. Most of the extra activity he found would reflect impulses propagating belatedly through a potentially dangerous reentrant channel. Put another way, Simson thought he could place electrodes on the skin, as for a standard electrocardiogram, but then record only those currents produced in the 40 milliseconds immediately after formation of the familiar QRS wave seen on electrocardiograms. (The QRS wave reflects the spread of impulses through the ventricles.) Heart cells are generally quiet at that point, giving rise to a flat line on the electrocardiogram tracing. Signal-averaged deviations from this normal pattern would signify slow conduction in a reentrant pathway.

Simson spent that night in his basement building a signal-averaging device. The next day Josephson, Horowitz and I were scheduled to remove tissue that had earlier caused reentrant arrhythmia in one of our patients. Before surgery, Simson attached his new recorder to the patient and noted, as expected, that there was a flurry of electrical activity in the usually quiescent span following ventricular excitation. But was the signal, in fact, an indication of late impulse conduction in a reentrant circuit? The answer would be yes if the fluctuations disappeared after the operation. The surgical procedure went well. Josephson and Horowitz identified the circuit, and I excised the entryway. After surgery, Simson reattached his device to the patient. The post-QRS fluctuations were gone.

We had come a long way since 1978. We had learned why our surgical approach, initially designed by guesswork, is useful. It interrupts the diseased anatomic pathway that, in response to aberrant firing by a nearby cell, gives rise to the repeated flow of impulses through a recursive circuit. Moreover, we had gained the ability to identify noninvasively patients at risk.

At the University of Colorado, where I moved in 1984, we use Simson's screening test routinely. We usually wait two or three months after a heart attack to be sure we are not detecting a predisposition to "automatic" tachycardias. For a week or so after a person has a heart attack, dying cells often fire when they should be silent. This behavior can cause the heart to beat prematurely. If the cell depolarizes repeatedly, the activity could lead to fast beating, and sometimes failure, of the heart. A tendency to automatic tachycardia generally resolves within a few weeks, as the sputtering cells expire.

If a propensity for reentrant tachycardia is discovered after a suitable waiting period, and if medications do not suffice, patients can consider other treatment options. I speak of more than one choice because surgery is no longer the only therapeutic alternative to drugs. A device known as an implantable defibrillator has been available since 1980. When the heart begins to beat quickly, the machine issues a shock that depolarizes the entire heart instantly, giving the sinoatrial node a chance to resume its pacemaker function.

About half as many patients die from complications of the implantation procedure for the device as from consequences of undergoing our surgery. But, in contrast to the surgery, the device offers only palliation, not a cure. Recipients continue to face episodes of tachycardia and may lose consciousness each time they are shocked back into normal rhythm. Consequently, they cannot drive or engage in other activities where sudden blackouts could be dangerous. If surgery to eliminate a reentrant circuit is deemed the better therapy for a given patient, it can now be obtained at many medical centers.

Overall, it is fair to say that the majority of patients who survive a heart attack are not vulnerable to reentrant arrhythmias. Perhaps half of the small group who are susceptible can be treated with medication. Of those who do not respond to drugs, however, as many as 80 percent are likely to die from their electrical abnormality within a year after their first bout of reentrant tachycardia unless they receive some other therapy. It is reassuring to know that for many of those individuals the courage of a Philadelphia banker has permitted a cure.

FURTHER READING

OBSERVATIONS ON MECHANISMS OF VENTRICULAR TACHYCARDIA IN MAN. H.J.J. Wellens, D. R. Duren and K. I. Lie in *Circulation,* Vol. 54, No. 2, pages 237–244; August 1976.

SURGICAL ENDOCARDIAL RESECTION FOR THE TREATMENT OF MALIGNANT VENTRICULAR TACHYCARDIA. A. H. Harken, M. E. Josephson and L. N. Horowitz in *Annals of Surgery,* Vol. 190, No. 4, pages 456–460; October 1979.

CARDIAC ARRHYTHMIAS. A. H. Harken in *Care of the Surgical Patient,* Vol. 1: *Critical Care.* Edited by D. W. Wilmore, M. F. Brennan, A. H. Harken, J. W. Holcroft and J. L. Meakins. Scientific American Medicine, 1992.

Fuzzy Logic

The binary logic of modern computers often falls short when describing the vagueness of the real world. Fuzzy logic offers more graceful alternatives

by Bart Kosko and Satoru Isaka

Computers do not reason as brains do. Computers "reason" when they manipulate precise facts that have been reduced to strings of zeros and ones and statements that are either true or false. The human brain can reason with vague assertions or claims that involve uncertainties or value judgments: "The air is cool," or "That speed is fast" or "She is young." Unlike computers, humans have common sense that enables them to reason in a world where things are only partially true.

Fuzzy logic is a branch of machine intelligence that helps computers paint gray, commonsense pictures of an uncertain world. Logicians in the 1920s first broached its key concept: everything is a matter of degree.

Fuzzy logic manipulates such vague concepts as "warm" or "still dirty" and so helps engineers to build air conditioners, washing machines and other devices that judge how fast they should operate or shift from one setting to another even when the criteria for making those changes are hard to define. When mathematicians lack specific algorithms that dictate how a system should respond to inputs, fuzzy logic can control or describe the system by using "commonsense" rules that refer to indefinite quantities. No known mathematical model can back up a truck-and-trailer rig from a parking lot to a loading dock when the vehicle starts from a random spot. Both humans and fuzzy systems can perform this nonlinear guidance task by using practical but imprecise rules such as "If the trailer turns a little to the left, then turn it a little to the right." Fuzzy systems often glean their rules from experts. When no expert gives the rules, adaptive fuzzy systems learn the rules by observing how people regulate real systems.

A recent wave of commercial fuzzy products, most of them from Japan, has popularized fuzzy logic. In 1980 the contracting firm of F. L. Smidth & Company in Copenhagen first used a fuzzy system to oversee the operation of a cement kiln. In 1988 Hitachi turned over control of a subway in Sendai, Japan, to a fuzzy system. Since then, Japanese companies have used fuzzy logic to direct hundreds of household appliances and electronics products. The Ministry of International Trade and Industry estimates that in 1992 Japan produced about $2 billion worth of fuzzy products. U.S. and European companies still lag far behind.

Applications for fuzzy logic extend beyond control systems. Recent theorems show that in principle fuzzy logic can be used to model any continuous system, be it based in engineering or physics or biology or economics. Investigators in many fields may find that fuzzy, commonsense models are more useful or accurate than are standard mathematical ones.

At the heart of the difference between classical and fuzzy logic is something Aristotle called the law of the excluded middle. In standard set theory, an object either does or does not belong to a set. There is no middle ground: the number five belongs fully to the set of odd numbers and not at all to the set of even numbers. In such bivalent sets, an object cannot belong to both a set and its complement set or to neither of the sets. This principle preserves the structure of logic and avoids the contradiction of an object that both is and is not a thing at the same time.

Sets that are fuzzy, or multivalent, break the law of the excluded middle—to some degree. Items belong only partially to a fuzzy set. They may also belong to more than one set. Even to just one individual, the air may feel cool, just right and warm to varying degrees. Whereas the boundaries of standard sets are exact, those of fuzzy sets are curved or taper off, and this curvature creates partial contradictions. The air can be 20 percent cool—and at the same time, 80 percent not cool.

Fuzzy degrees are not the same as probability percentages, a point that has eluded some critics of the field. Probabilities measure whether something will occur or not. Fuzziness measures the degree to which something occurs or some condition exists. The statement "There is a 30 percent chance the weather will be cool" conveys the probability of cool weather. But "The morning feels 30 percent cool" means that the air feels cool to some extent—and at the same time, just right and warm to varying extents.

The only constraint on fuzzy logic is that an object's degrees of membership in complementary groups must sum to unity. If the air seems 20 percent cool, it must also be 80 percent not cool. In this way, fuzzy logic just skirts the bivalent contradiction—that something is 100 percent cool and 100 percent not cool—that would destroy formal logic. The law of the excluded middle holds merely as a special case in fuzzy logic, namely when an object belongs 100 percent to one group.

The modern study of fuzzy logic and partial contradictions had its origins early in this century, when Bertrand Russell found the ancient Greek paradox at

BART KOSKO and SATORU ISAKA are pioneers in the development of fuzzy logic systems. Kosko holds degrees in philosophy and economics from the University of Southern California, a master's in applied mathematics from the University of California, San Diego, and a Ph.D. in electrical engineering from the University of California, Irvine. He is an assistant professor of electrical engineering at U.S.C., a governor of the International Neural Network Society and the program co-chair of the 1993 World Congress on Neural Networks. Isaka specializes in fuzzy information processing in the research and development division at Omron Advanced Systems in Santa Clara, Calif. He is interested in applying machine learning and adaptive control systems to biomedical systems and factory automation. He received his M.S. and Ph.D. degrees in systems science from U.C.S.D., in 1986 and 1989, respectively.

the core of modern set theory and logic. According to the old riddle, a Cretan asserts that all Cretans lie. So, is he lying? If he lies, then he tells the truth and does not lie. If he does not lie, then he tells the truth and so lies. Both cases lead to a contradiction because the statement is both true and false. Russell found the same paradox in set theory. The set of all sets is a set, and so it is a member of itself. Yet the set of all apples is not a member of itself because its members are apples and not sets. Perceiving the underlying contradiction, Russell then asked, "Is the set of all sets that are not members of themselves a member of itself?" If it is, it isn't; if it isn't, it is.

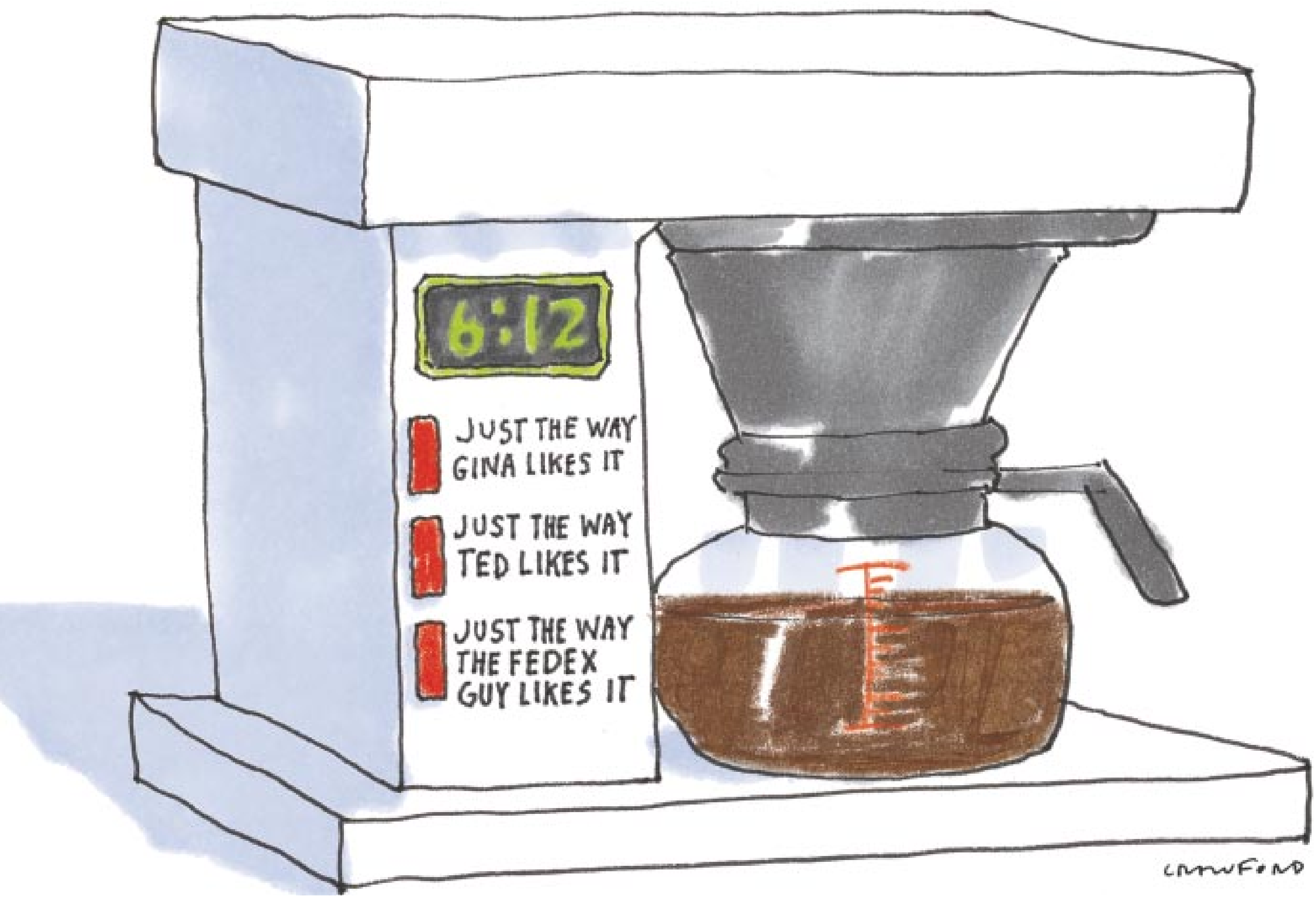
Faced with such a conundrum, classical logic surrenders. But fuzzy logic says that the answer is half true and half false, a 50–50 divide. Fifty percent of the Cretan's statements are true, and 50 percent are false. The Cretan lies 50 percent of the time and does not lie the other half. When membership is less than total, a bivalent system might simplify the problem by rounding it down to zero or up to 100 percent. Yet 50 percent does not round up or down.

In the 1920s, independent of Russell, the Polish logician Jan Łukasiewicz worked out the principles of multivalued logic, in which statements can take on fractional truth values between the ones and zeros of binary logic. In a 1937 article in *Philosophy of Science,* quantum philosopher Max Black applied multivalued logic to lists, or sets of objects, and in so doing drew the first fuzzy set curves. Following Russell's lead, Black called the sets "vague."

Almost 30 years later Lotfi A. Zadeh, then chair of the electrical engineering department at the University of California at Berkeley, published "Fuzzy Sets," a landmark paper that gave the field its name. Zadeh applied Łukasiewicz's logic to every object in a set and worked out a complete algebra for fuzzy sets. Even so, fuzzy sets were not put to use until the mid-1970s, when Ebrahim H. Mamdani of Queen Mary College in London designed a fuzzy controller for a steam engine. Since then, the term "fuzzy logic" has come to mean any mathematical or computer system that reasons with fuzzy sets.

Fuzzy logic is based on rules of the form "if...then" that convert inputs to outputs—one fuzzy set into another. The controller of a car's air conditioner might include rules such as "If the temperature is cool, then set the motor speed on slow" and "If the temperature is just right, then set the motor speed on medium." The temperatures (cool, just right) and the motor speeds (slow, medium) name fuzzy sets rather than specific values.

To build a fuzzy system, an engineer might begin with a set of fuzzy rules from an expert. An engineer might define the degrees of membership in various fuzzy input and output sets with sets of curves. The relation between the input and output sets could then be plotted. Given the rule "If the air feels cool, then set the motor to slow," the inputs (temperature) would be listed along



BUILDING MACHINES THAT UNDERSTAND the vagueness of the real world is the goal of those working on fuzzy logic. Fuzzy logic can model and control nuances overlooked by the binary logic of conventional computers.

one axis of a graph and the outputs (motor speed) along a second axis. The product of these fuzzy sets forms a fuzzy patch, an area that represents the set of all associations that the rule forms between those inputs and outputs.

The size of the patch reflects the rule's vagueness or uncertainty. The more precise the fuzzy set, the smaller it becomes. If "cool" is precisely 68 degrees Fahrenheit, the fuzzy set collapses to a spike. If both the cool and slow fuzzy sets are spikes, the rule patch is a point.

The rules of a fuzzy system define a set of overlapping patches that relate a full range of inputs to a full range of outputs. In that sense, the fuzzy system approximates some mathematical function or equation of cause and effect. These functions might be laws that tell a microprocessor how to adjust the power of an air conditioner or the speed of a washing machine in response to some fresh measurement.

Fuzzy systems can approximate any continuous math function. One of us (Kosko) proved this uniform convergence theorem by showing that enough small fuzzy patches can sufficiently cover the graph of any function or input/output relation. The theorem also shows that we can pick in advance the maximum error of the approximation and be sure there is a finite number of fuzzy rules that achieve it. A fuzzy system reasons, or infers, based on its rule patches. Two or more rules convert any incoming number into some result because the patches overlap. When data trigger the rules, overlapping patches fire in parallel—but only to some degree.

Imagine a fuzzy air conditioner that relies on five rules and thus five patches to match temperatures to motor speeds. The temperature sets (cold, cool, just right, warm and hot) cover all the possible fuzzy inputs. The motor speed sets (very slow, slow, medium, fast and very fast) describe all the fuzzy outputs. A temperature of, say, 68 degrees F might be 20 percent cool (80 percent not cool) and 70 percent just right (30 percent not just right). At the same time, the air is also 0 percent cold, warm and hot. The "if cool" and "if just right" rules would fire and invoke both the slow and medium motor speeds.

The two rules contribute proportionally to the final motor speed. Because the temperature was 20 percent cool, the curve describing the slow engine speed must shrink to 20 percent of its height. The "medium" curve must shrink to 70 percent. Summing those two reduced curves produces the final curve for the fuzzy output set.

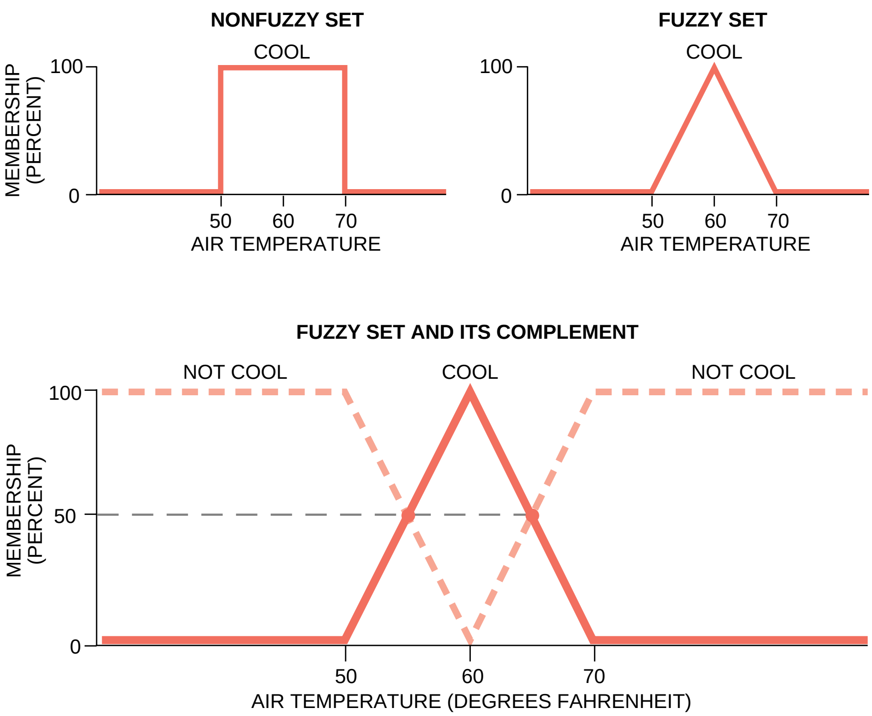
In its fuzzy form, such an output curve does not assist controllers that act on binary instructions. So the final step is a process of defuzzification, in which the fuzzy output curve is turned into a single numerical value. The most common technique is to compute the center of mass, or centroid, of the area under the curve. In this instance, the centroid of the fuzzy output curve might correspond to a motor speed of 47 revolutions per minute. Thus, beginning with a quantitative temperature input, the electronic controller can reason from fuzzy temperature and motor speed sets and arrive at an appropriate and precise speed output.

All fuzzy systems reason with this fire-and-sum technique—or something close to it. As systems become more complex, the antecedents of the rules may include any number of terms conjoined by "and" or disjoined by "or." An advanced fuzzy air conditioner might use a rule that says, "If the air is cool and the humidity is high, then set the motor to medium."

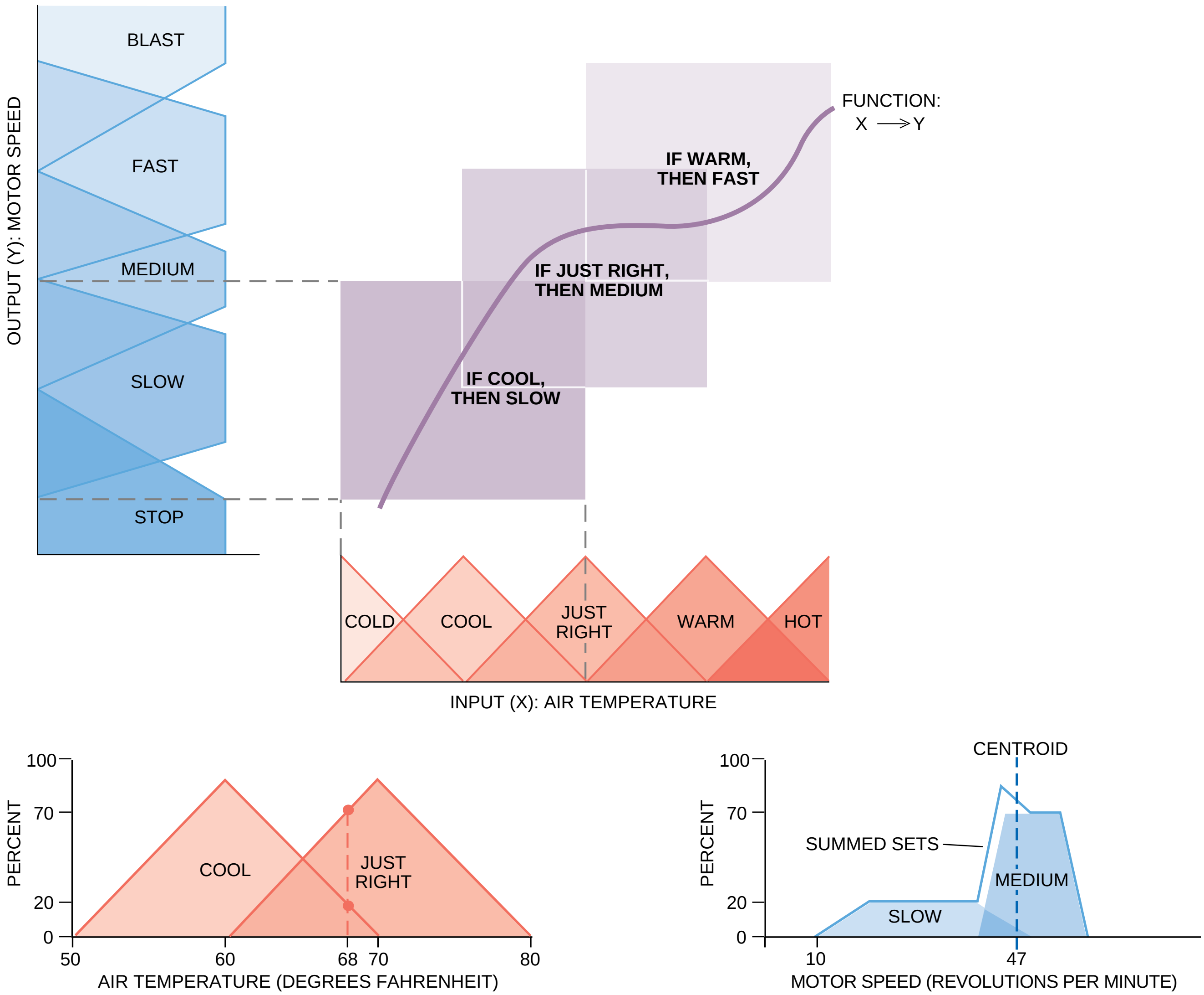
Fuzzy products use both microprocessors that run fuzzy inference algorithms and sensors that measure changing input conditions. Fuzzy chips are microprocessors designed to store and process fuzzy rules. In 1985 Masaki Togai and Hiroyuki Watanabe, then working at AT&T Bell Laboratories, built the first digital fuzzy chip. It processed 16 simple rules in 12.5 microseconds, a rate of 0.08 million fuzzy logical inferences per second. Togai InfraLogic, Inc., now offers chips based on Fuzzy Computational Acceleration hardware that processes up to two million rules per second. Most microprocessor firms currently have fuzzy chip research projects. Fuzzy products largely rely on standard microprocessors that engineers have programmed with a few lines of fuzzy inference code. Although the market for dedicated fuzzy chips is still tiny, the value of microprocessors that include fuzzy logic already exceeds $1 billion.

The most famous fuzzy application is the subway car controller used in Sendai, which has outperformed both human operators and conventional automated controllers. Conventional controllers start or stop a train by reacting to position markers that show how far the vehicle is from a station. Because the controllers are rigidly programmed, the ride may be jerky: the automated controller will apply the same brake pressure when a train is, say, 100 meters from a station, even if the train is going uphill or downhill.

In the mid-1980s engineers from Hitachi used fuzzy rules to accelerate, slow and brake the subway trains more smoothly than could a deft human operator. The rules encompassed a broad



SET THEORY underlies the difference between standard and fuzzy logic. In standard logic, objects belong to a set fully or not at all (*top left*). Objects belong to a fuzzy set only to some extent (*top right*) and to the set's complement to some extent. Those partial memberships must sum to unity (*bottom*). If 55 degrees is 50 percent "cool," it is also 50 percent "not cool."



APPLICATION OF FUZZY LOGIC to the control of an air conditioner shows how manipulating vague sets can yield precise instructions. The air conditioner measures air temperature and then calculates the appropriate motor speed. The system uses rules that associate fuzzy sets of temperatures, such as "cool," to fuzzy sets of motor outputs, such as "slow." Each rule forms a fuzzy patch. A chain of patches can approximate a performance curve or other function (*top*). If a temperature of 68 degrees Fahrenheit is 20 percent "cool" and 70 percent "just right" (*bottom left*), two rules fire, and the system tries to run its motor at a speed that is 20 percent "slow" and 70 percent "medium" (*bottom right*). The system arrives at an exact motor speed by finding the center of mass, or centroid, for the sum of the motor output curves.

range of variables about the ongoing performance of the train, such as how frequently and by how much its speed changed and how close the actual speed was to the maximum speed. In simulated tests the fuzzy controller beat an automated version on measures of riders' comfort, shortened riding times and even achieved a 10 percent reduction in the train's energy consumption. Today the fuzzy system runs the Sendai subway during peak hours and runs some Tokyo trains as well. Humans operate the subway during nonpeak hours to keep up their skills.

Companies in Japan and Korea are building an array of fuzzy consumer goods that offer more precise control than do conventional ones. Fuzzy washing machines adjust the wash cycle to every set of clothes, changing strategies as the clothes become clean. A fuzzy washing machine gives a finer wash than a "dumb" machine with fixed commands. In the simplest of these machines, an optical sensor measures the murk or clarity of the wash water, and the controller estimates how long it would take a stain to dissolve or saturate in the wash water. Some machines use a load sensor to trigger changes in the agitation rate or water temperature. Others shoot bubbles into the wash to help dissolve dirt and detergent. A washing machine may use as few as 10 fuzzy rules to determine a wide variety of washing strategies.

In cameras and camcorders, fuzzy logic links image data to various lens settings. One of the first fuzzy camcorders, the Canon hand-held H800, which was introduced in 1990, adjusts the autofocus based on 13 fuzzy rules. Sensors measure the clarity of images in six areas. The rules take up about a kilobyte of memory and convert the sensor data to new lens settings.

Matsushita relies on more rules to cancel the image jitter that a shaking hand causes in its small Panasonic camcorders. The fuzzy rules infer where the image will shift. The rules heed local and global changes in the image and then compensate for them. In contrast, camcorder controllers based on mathematical models can compensate for no more than a few types of image jitter.

Systems with fuzzy controllers are often more energy efficient because they calculate more precisely how much power is needed to get a job done. Mitsubishi and Korea's Samsung report that their fuzzy vacuum cleaners achieve more than 40 percent energy savings over nonfuzzy designs. The fuzzy systems use infrared light-emitting diodes to measure changes in dust flow and so to judge if a floor is bare. A four-bit microprocessor measures the dust flow to calculate the appropriate suction power and other vacuum settings.

Automobiles also benefit from fuzzy logic. General Motors uses a fuzzy transmission in its Saturn. Nissan has patented a fuzzy antiskid braking system, fuzzy transmission system and fuzzy fuel injector. One set of fuzzy rules in an on-board microprocessor adjusts the fuel flow. Sensors measure the throttle setting, manifold pressure, radiator water temperature and the engine's revolutions per minute. A second set of fuzzy rules times the engine ignition based on the revolutions per minute, water temperature and oxygen concentration.

One of the most complex fuzzy systems is a model helicopter, designed by Michio Sugeno of the Tokyo Institute of Technology. Four elements of the craft—the elevator, aileron, throttle and rudder—respond to 13 fuzzy voice commands, such as "up," "land" and "hover." The fuzzy controller can make the craft hover in place, a difficult task even for human pilots.

A few fuzzy systems manage information rather than devices. With fuzzy logic rules, the Japanese conglomerate Omron oversees five medical data bases in a health management system for large firms. The fuzzy systems use 500 rules to diagnose the health of some 10,000 patients and to draw up personalized plans to help them prevent disease, stay fit and reduce stress. Other companies, including Hitachi and Yamaichi Securities, have built trading programs for bonds or stock funds that use fuzzy rules to react to changes in economic data.

The Achilles' heel of a fuzzy system is its rules. Almost all the fuzzy consumer products now on the market rely on rules supplied by an expert. Engineers then engage in a lengthy process of tuning those rules and the fuzzy sets. To automate this process, some engineers are building adaptive fuzzy systems that use neural networks or other statistical tools to refine or even form those initial rules.

Neural networks are collections of "neurons" and "synapses" that change their values in response to inputs from surrounding neurons and synapses. The neural net acts like a computer because it maps inputs to outputs. The neurons and synapses may be silicon components or equations in software that simulate their behavior. A neuron adds up all the incoming signals from other neurons and then emits its own response in the form of a number. Signals travel across the synapses, which have numerical values that weight the flow of neuronal signals. When new input data fire a network's neurons, the synaptic values can change slightly. A neural net "learns" when it changes the value of its synapses.

Depending on the available data, networks can learn patterns with or without supervision. A supervised net learns by trial and error, guided by a teacher. A human may point out when the network has erred—when it has emitted a response that differs from the desired output. The teacher will correct the responses to sample data until the network responds correctly to every input.

Supervised networks tune the rules of a fuzzy system as if they were synapses. The user provides the first set of rules, which the neural net refines by running through hundreds of thousands of inputs, slightly varying the fuzzy sets each time to see how well the system performs. The net tends to keep the changes that improve performance and to ignore the others.

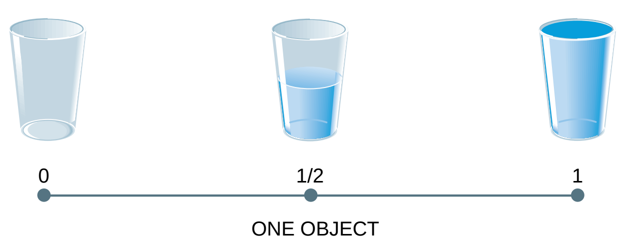
A handful of products in Japan now use supervised neural learning to tune the fuzzy rules that control their operation. Among them are Sanyo's microwave oven and several companies' washing machines. Sharp employs this technique to modify the rules of its fuzzy refrigerator so that the device learns how often its hungry patron is likely to open the door and adjusts the cooling cycle accordingly. So far the neural net must learn "off-line" in the laboratory, from small samples of behavior by average customers. In time, researchers at such groups as Japan's Laboratory for International Fuzzy Engineering and the Fuzzy Logic Systems Institute hope to build fuzzy systems that will adapt to the needs of each consumer.

Supervised networks do have drawbacks. Tuning such systems can take hours or days of computer time be-

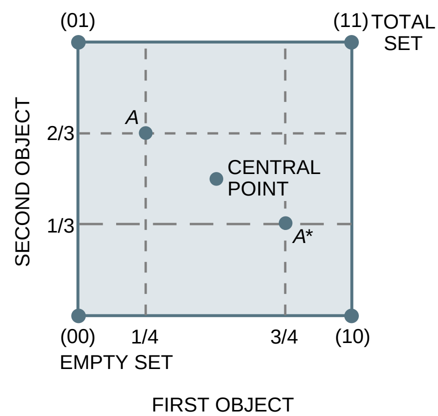
Fuzzy Sets and Paradoxes

Relations between sets show the paradox at the heart of fuzzy logic. In standard sets, an object either does or does not belong to a set: the glass of water is either empty or full. Fuzzy sets cover a continuum of partial sets, such as glasses only half full (*left*).

Two objects or two sets define a two-dimensional space (*right*). Standard set theory holds only at the corners of the square, where the values are either 1 or 0. The midpoint, the fuzziest point in the square, cannot be rounded off to a corner.

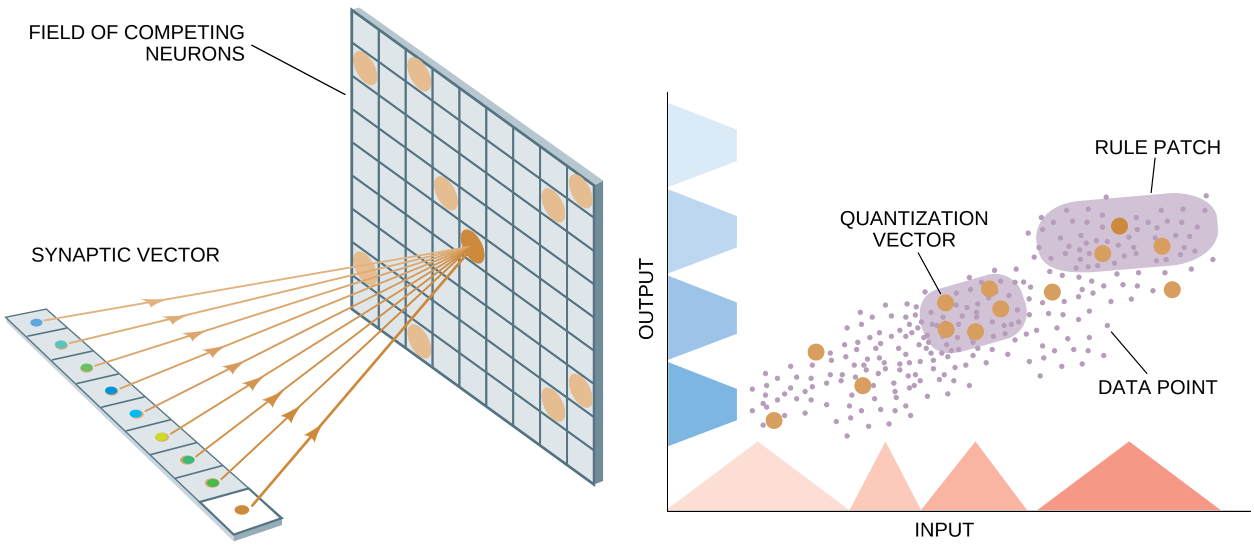


Fuzzy sets of two objects are points of the square, such as the point *A*. The union and intersection of a fuzzy set *A* and its complement *A** lie at the corners of an interior square. In contrast, the union or intersection of a standard set and its complement results in the total set or the empty set at the square's corners, respectively.



How Fuzzy Systems Learn Rules

Adaptive systems called neural networks can help fuzzy systems learn rules. A neural network accepts pairs of input and output data, such as temperatures and motor speeds for air conditioners, and groups them into a small number of prototypes, or classes. Within the network, each prototype acts as a quantization vector—a list of numbers—that stands for the synapses feeding into a neuron. When a new data point enters the network, it stimulates the neuron associated with the prototype that matches the data most closely. The values of the "winning" synapses adjust to reflect the data they are receiving. As the data cluster, so do the quantization vectors, which define rule patches. More data lead to more numerous and precise patches.



cause networks may converge on an inappropriate solution or rule or may fail to converge at all. Neural researchers have proposed hundreds of schemes to alleviate this problem, but none has removed it. Even after a lengthy tuning session, the final rules may not be much better than the first set.

Rather than relying on an expert to supply a training set of data and correct a network in the process of learning, unsupervised neural networks learn simply by observing an expert's decisions. In this way, an adaptive fuzzy system can learn to spot rule patterns in the incoming data. Broad rule patches form quickly, based on a few inputs. Those patches are refined over time.

Unsupervised neural networks blindly cluster data into groups, the members of which resemble one another. There may be no given right or wrong response or way to organize the data. The algorithms are simpler, and, at least in theory, the network need run through the data just once. (In some cases, when data are sparse, the neural net must cycle through them repeatedly.) Unsupervised learning is thus much faster than supervised learning. With numerical inputs and outputs supplied by an expert or a physical process or even an algorithm, an unsupervised neural network can find the first set of rules for a fuzzy system. The quality of the rules depends on the quality of the data and therefore on the skills of the expert who generates the data. At this point, there are fewer unsupervised than supervised adaptive fuzzy systems. Because unsupervised networks are best used to create rules and supervised networks are better at refining them, hybrid adaptive fuzzy systems include both.

Most fuzzy systems have been control systems with few variables. That trend happened because most of the first fuzzy logic engineers were control theorists and because a control loop regulates most consumer products. The challenge for the next generation of fuzzy research will be tackling large-scale, nonlinear systems with many variables. These problems can arise when people try to supervise manufacturing plants or schedule airline flights or model the economy. No experts may be able to describe such systems. Common sense may falter or not apply. The neural nets that must learn the rules for modeling these hard problems may have little or no data to go on.

A further problem is that, like any other mathematical or computer model, fuzzy logic falls prey to the "curse of dimensionality": the number of fuzzy rules tends to grow exponentially as the number of system variables increases. Fuzzy systems must contend with a trade-off. Large rule patches mean the system is more manageable but also less precise. Even with that trade-off, fuzzy logic can often better model the vagueness of the world than can the black-and-white concepts of set theory. For that reason, fuzzy logic systems may well find their way into an ever growing number of computers, home appliances and theoretical models. The next century may be fuzzier than we think.

FURTHER READING

VAGUENESS: AN EXERCISE IN LOGICAL ANALYSIS. Max Black in *Philosophy of Science*, Vol. 4, No. 4, pages 427–455; October 1937.

FUZZY SETS AND APPLICATIONS: SELECTED PAPERS BY L. A. ZADEH. Edited by R. R. Yager et al. Wiley-Interscience, 1987.

NEURAL NETWORKS AND FUZZY SYSTEMS. Bart Kosko. Prentice-Hall, 1991.

PROGRESS IN SUPERVISED NEURAL NETWORKS. Don R. Hush and Bill G. Horne in *IEEE Signal Processing Magazine*, Vol. 10, No. 1, pages 8–39; January 1993.

Edwin Hubble and the Expanding Universe

More than any other individual, he shaped astronomers' present understanding of an expanding universe populated by a multitude of galaxies

by Donald E. Osterbrock, Joel A. Gwinn and Ronald S. Brashear

During the 1920s and 1930s, Edwin Powell Hubble changed the scientific understanding of the universe more profoundly than had any astronomer since Galileo. Much as Galileo banished the earth from the center of the solar system, Hubble proved that, rather than being unique, the Milky Way is but one of untold millions of galaxies, or "island universes." Hubble's work also helped to replace the notion of a static cosmos with the startling view that the entire universe is expanding, the ultimate extension of Galileo's defiant (if apocryphal) assertion, "Yet still it moves." Although many researchers contributed to those revolutionary discoveries, Hubble's energetic drive, keen intellect and supple communication skills enabled him to seize the problem of the construction of the universe and make it peculiarly his own.

Hubble's early years have become enmeshed in myth, in part because of his desire to play the hero and in part because of the romanticized image of him recorded by his wife, Grace, in her journals. Many accounts of Hubble consequently bear little relation to the real man. For example, his biographers often state that he was a professional-caliber boxer, yet there is no evidence to support that claim. Likewise, the oft-repeated story that he was wounded during service in World War I seems to have no basis in fact.

Even without such embellishments, Hubble's life has the ring of an all-American success story. The Hubble family established itself in Missouri in 1833, when Edwin's great-grandfather rode horseback from Virginia to settle in what is now Boone County. Edwin's grandfather and his father, John P. Hubble, were hard-working, moderately prosperous insurance agents. His mother, Virginia Lee James Hubble, was ninth in direct descent from Miles Standish. With characteristic overexuberance, Grace Hubble described her husband's ancestors thus: "Tall, well-made, strong, their bodily inheritance had come down to him, even to the clear, smooth skin that tanned in the sun, and the brown hair with a glint of reddish gold. They had handed down their traditions as well, integrity, loyalty as citizens, loyalty to their families . . . and a sturdy reliance on their own efforts."

Edwin was born in Marshfield, Mo., on November 20, 1889, the third of seven children who survived. In 1898 the Hubbles moved to Evanston, Ill., and, two years later, to Wheaton, where Edwin showed a robust aptitude for both academics and athletics. He became a star athlete at Wheaton High School, especially on the track team. During those years, Edwin also consistently placed in the top quarter of his class. At the age of 16, he entered the University of Chicago, where he continued to display his twin talents. He earned high grades in mathematics, chemistry, physics, astronomy and languages. Despite being two years younger than most of his classmates, he won letters in track and basketball.

After graduation, Hubble received a Rhodes scholarship. Even then, it was far from clear what direction his wide-ranging abilities would finally take. When he "went up" to Queen's College at the University of Oxford in October 1910, Hubble "read" (studied) jurisprudence; both his father and grandfather hoped he would become a lawyer. After completing the course in two years with "second-class honors," he switched to Spanish in his third year. He also continued to participate actively in sports.

Edwin remained a Rhodes scholar for three years and then rejoined his family, who had since moved to Kentucky. He claimed to have passed the bar and practiced law in Louisville. According to Hubble's friend Nicholas U. Mayall, he "chucked the law for astronomy, and I knew that even if I were second-rate or third-rate, it was astronomy that mattered." No record of Hubble's bar exam or of his law practice exists, however. Rather, during this time he taught and coached basketball at the high school in New Albany, Ind., just across the Ohio River from Louisville.

Uninterested in the law and unsatisfied by his life as a teacher, Hubble decided to return to his true passion, astronomy. In the spring of 1914 Hubble wrote Forest Ray Moulton, his astronomy professor at Chicago, to ask about the possibility of returning as a graduate student. Moulton enthusiastically recommended him to Edwin B. Frost, the director of Yerkes Observatory in Williams Bay, Wis., where the graduate program was given.

Frost gladly accepted Hubble as a student and gave him a scholarship covering tuition and living expenses. The American Astronomical Society was to meet at Northwestern University in nearby Evanston in August 1914; Frost suggested that Hubble come north in time

DONALD E. OSTERBROCK, JOEL A. GWINN and RONALD S. BRASHEAR share a keen interest in Hubble's place in the history of astronomy. Osterbrock is a former director of Lick Observatory at Mount Hamilton, Calif., where he now researches and teaches astronomy. He has written many books and is currently working on a biography of George Willis Ritchey, a noted American telescope maker and designer. Gwinn has spent more than 30 years at the University of Louisville, where he is a professor of physics; he has concentrated on optics, atomic spectroscopy and the history of astronomy, including Hubble's early life. Brashear earned his master's degree in 1984 from Louisville, where he met Gwinn. Brashear is curator of the history of science at the Henry E. Huntington Library in San Marino, Calif.

EDWIN HUBBLE drew on his observing experience, personal drive and access to top facilities to make a series of groundbreaking cosmological discoveries. He is seen here in the observing cage of the 200-inch Hale telescope, circa 1950.

for the event. Thus it happened that Hubble was present at the meeting when Vesto M. Slipher, a quiet, modest astronomer at Lowell Observatory, fueled a controversy as a result of his latest studies of nebulae.

"Nebulae" was a blanket term used by astronomers for centuries to designate faint, cloudy objects that, unlike comets, do not change in position or appearance. The nature of these objects defied easy explanation. In 1755 Immanuel Kant had postulated that some nebulae might be so-called island universes—self-contained systems of stars like our own Milky Way. That notion excited many scientists' imagination but could not be easily proved.

During the 19th century, improved telescopic observations showed that many nebulae definitely consist of clouds of luminous gas. One noteworthy class of spiral-shaped nebulae looks distinctly unlike the others, however. By the beginning of the 20th century, many astronomers had come to believe that spiral nebulae were in fact distant galaxies composed of a multitude of stars; skeptics continued to argue that these objects were nearby structures, possibly infant stars in the process of formation.

At the 1914 Astronomical Society meeting, Slipher personally presented the first well-exposed, well-calibrated photographs of the spectra of spiral nebulae. Those photographs displayed light





HUBBLE'S FAMILY helped to instill in him an appreciation of self-motivation and hard work. Here he is seen with his cousins; Hubble appears third from the left (*top left*). At the time of his graduation from the University of Chicago (*top middle*), Hubble considered a career in law. This pose of Hubble with his sister Lucy, a Red Cross nurse, captures his sense of pride in uniform (*top right*). Hubble often went on weekend outings, such as this one with his high school students and their family members (*bottom left*). Late in life, he came to enjoy the solitary pleasures of fly-fishing (*bottom right*).

from the nebulae after it had been split into its constituent rainbow of colors. The spectra of stars contain dark lines, called absorption lines, caused by radiation captured by atoms in the stellar atmospheres. Slipher showed that the spectra of spiral nebulae exhibit the kinds of absorption-line spectra characteristic of collections of stars.

Moreover, Slipher found that the wavelengths of the nebular absorption lines are significantly offset from where they appear in the laboratory. Such changes in wavelength, or Doppler shifts, are caused by the motion of the object emitting the light toward or away from the earth. The Doppler shifts indicated that the spiral nebulae observed by Slipher are moving far faster than is typical for stars within the Milky Way and hence strengthened the argument that the spiral nebulae are not part of our galaxy.

When Hubble arrived at Yerkes Observatory, he found a near-moribund institution engaging in routine research. Undaunted, Hubble set out in his own direction. Possibly inspired by Slipher's presentation, he began an intriguing program of photographing nebulae using the observatory's underused 24-inch reflecting telescope. That work grew into his Ph.D. thesis, "Photographic Investigations of Faint Nebulae," which contains foreshadowings of his later research on galaxies and on cosmology.

Hubble described and classified the numerous small, faint nebulae that did not resemble diffuse clouds of gas and noted that most of those nebulae are not spiral but elliptical. He emphasized that the distribution of faint nebulae avoids the parts of the sky near the Milky Way and that many of them appear clustered together. Hubble left no doubt as to the direction of his thinking. "Suppose them to be extra-sidereal and perhaps we see clusters of galaxies; suppose them within our system, their nature becomes a mystery," he wrote. Hubble's thesis is technically sloppy and confused in its theoretical ideas, but it clearly shows the hand of a great scientist groping toward the solution of a profound problem.

In October 1916 Hubble corresponded with George Ellery Hale, the director of Mount Wilson Observatory in Pasadena, Calif. Hale was scouting for staff members to work with the 100-inch reflecting telescope nearing completion on Mount Wilson. Hale offered Hubble a job, conditional on the completion of his doctorate. Hubble hoped to finish the following June and then take the job at Mount Wilson.

Those plans were scuttled by the entry of the U.S. into World War I on April 6, 1917. Hubble, a former Rhodes scholar who had deep emotional ties to Great Britain, decided that patriotic duty took precedence over scientific pursuits. On May 15, just three days after receiving his degree, he reported for duty and began an officer training course.

Hubble thrived in the military setting. He became a battalion commander and next a major in the 86th "Black Hawk" Division. In September 1918 the division embarked for Europe, but, much to Hubble's regret, the Armistice arrived before he went into combat. He summed up his military career to Frost in these words: "I am still a major. I barely got under fire and altogather [sic] I am disappointed in the matter of the War. However, that Chapter is finished and another is opening."

Hubble was discharged on August 20, 1919, whereupon he immediately went to Mount Wilson. He had arrived at an opportune time. The observatory now had two huge reflecting telescopes, a 60-inch and the new 100-inch, then the largest in the world. Hubble's studies at Yerkes had taught him how to make efficient use of such telescopes, and his Ph.D. research had imparted a firm direction to his investigations.

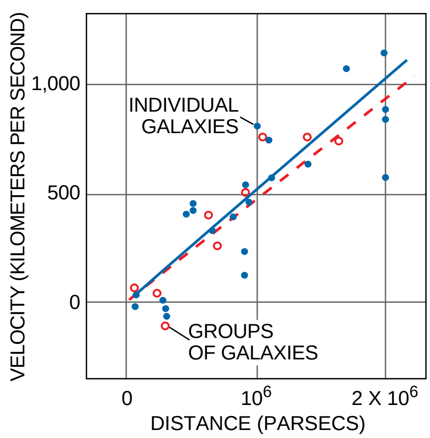
Milton L. Humason, who later collaborated with Hubble on his studies of the distant universe, vividly recalled Hubble's first observing session at Mount Wilson: "'Seeing' that night was rated as extremely poor on our Mount Wilson scale, but when Hubble came back from developing his plate he was jubilant. 'If this is a sample of poor seeing conditions,' he said, 'I shall always be able to get usable photographs with the Mount Wilson instruments.' The confidence and enthusiasm which he showed on that night were typical of the way he approached all his problems. He was sure of himself—of what he wanted to do, and of how to do it."

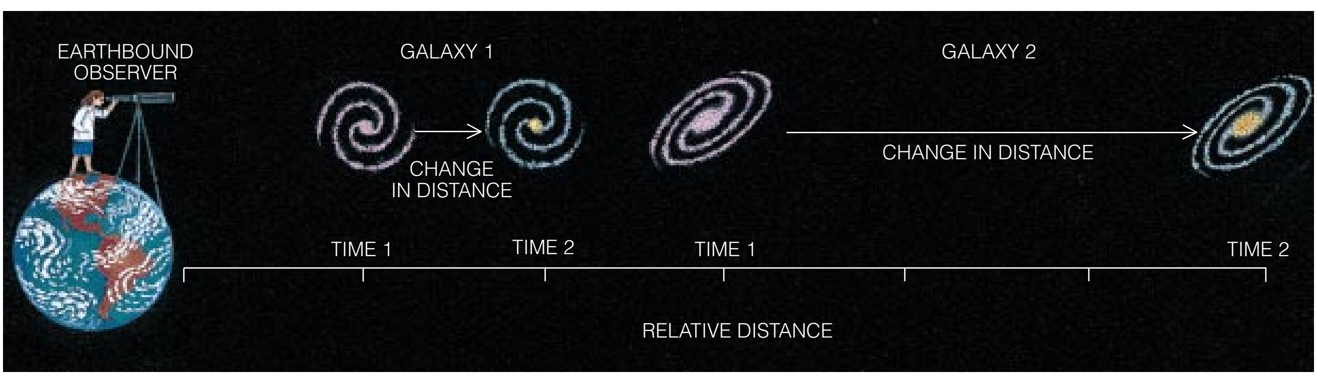
Hubble promptly returned to the question of the nature of the "nongalactic nebulae," those whose "members tend to avoid the galactic plane and concentrate in high galactic latitudes." At the time, he had not yet fully accepted the notion that these objects are galaxies outside our own. He initiated his detailed study of nongalactic nebulae by concentrating on the irregular object NGC 6822 ("NGC" denotes J.L.E. Dreyer's *New General Catalog* of nebulae). By 1923, using the 100-inch telescope, he had found several small nebulae and 12 variable stars within NGC 6822.

This endeavor suffered a happy interruption in 1924, when Hubble married Grace Burke Leib, a widow whose first husband had died in an accident in a coal mine in 1921. Grace had met Edwin during a trip to Mount Wilson in 1920 and idolized him for the rest of her life. Years later she recalled her electrifying first impression of Edwin as "an Olympian, tall, strong and beautiful, with the shoulders of the Hermes of Praxiteles, and the benign serenity."

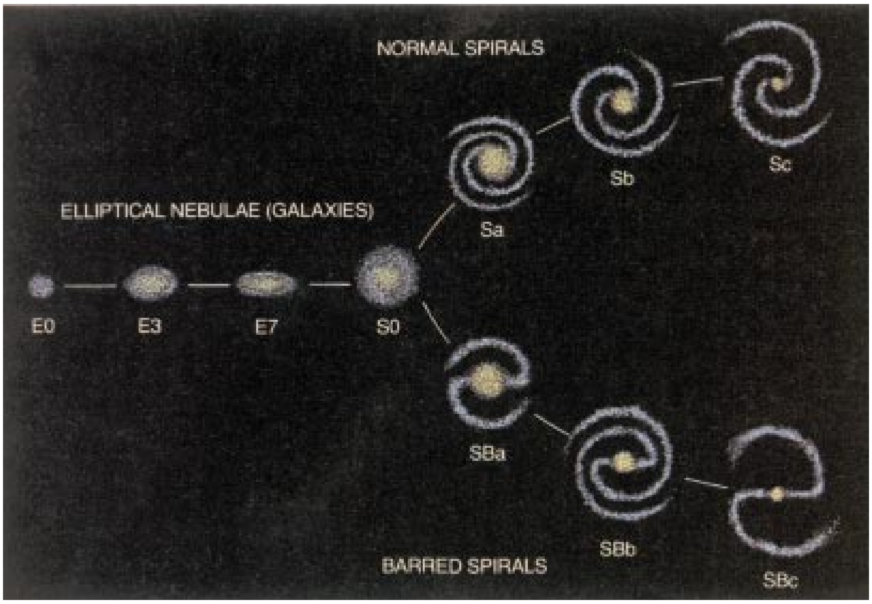
The newlyweds honeymooned in Carmel and in Europe. Hubble then returned to Pasadena, where he turned his attention to M31, the famous Andromeda galaxy. Drawing on his well-honed observing skills and the unequaled capabilities of the giant 100-inch telescope, Hubble managed to resolve distinctly six variable stars in M31. To Hubble's mind, the achievement unequivocally confirmed that M31 is a remote system composed of stars and, by implication, that other, fainter spiral nebulae are even more distant galaxies.

Although the brightest star that Hubble detected was only about 18th magnitude at its maximum—60,000 times fainter than the dimmest stars visible to the naked eye—he managed to make 83 measurements of its brightness, an impressive feat at the time. From these data, he determined that the star's brightness rose and fell in the manner characteristic of a class of stars known as Cepheid variables. These stars were of tremendous interest because the period of variation of a Cepheid directly relates to the star's absolute luminosity. By comparing that luminosity with his observations of the star's apparent brightness, Hubble was able to deduce





HUBBLE'S LAW states that galaxies appear to be receding at speeds in direct proportion to their distance from the observer. A graph of distance versus velocity (*top*), which Hubble published in 1929, plainly shows the linear relation. That relation is caused by the fact that the universe is expanding uniformly in all directions. The schematic illustration (*bottom*) shows why that is so. Galaxy 2 is three times as faraway as galaxy 1, so it has moved three times as far during the duration from time 1 to time 2. Earthbound observers therefore see galaxy 2 receding at three times the rate of galaxy 1.



GALAXY CLASSIFICATION scheme developed by Hubble in the 1920s assumed that elliptical galaxies develop arms and turn into normal or barred spiral galaxies. Astronomers have rejected Hubble's evolutionary theory but still use his diagram and nomenclature to categorize the wide variety of galaxy shapes.

the distance to the star and its surrounding galaxy.

A demon of energy, Hubble made discoveries in rapid succession. By the end of 1924 he had identified several more Cepheid variables in M31 and in M33, a spiral galaxy in the constellation Triangulum. He sent a paper describing his results to the joint American Astronomical Society–American Association for the Advancement of Science meeting in Washington, D.C., where the eminent astrophysicist Henry Norris Russell read the paper for him on January 1, 1925. Hubble had deduced that both M31 and M33 are about 930,000 light-years distant, far beyond the known bounds of the Milky Way. In the words of Joel Stebbins, the astronomical society's secretary, Hubble had succeeded at last in "bringing confirmation to the so-called island universe theory."

In contrast to the rapid progress of his research, Hubble published his results in a leisurely fashion, trying to be as certain as possible before committing them to print. The first paper containing Hubble's results from his studies of variable stars in spiral nebulae appeared later in 1925. Based on observations of Cepheids in NGC 6822, he declared it to be "the first object definitely assigned to a region outside the galactic system." Hubble's epochal paper, "A Spiral Nebula as a Stellar System," presenting his evidence that M33 also is a galaxy lying far outside our own, appeared in 1926. Not until 1929 did Hubble wrap up this phase of his research in a long paper that finally laid out in full detail his studies of the distance to and stellar composition of M31.

In addition to determining the distance to the "extra-galactic nebulae," Hubble sought to classify them and to comprehend their diversity of structure. Starting in 1923, he ordered the nebulae according to a perceived pattern of evolution. As an elliptical nebula ages, Hubble argued, it forms arms that enlarge and open up. To explain the existence of barred spirals, in which the spiral arms trail from a barlike formation, Hubble suggested that the elliptical could evolve into either a normal spiral or the barred form. The resulting double-pronged classification diagram is still widely used, though stripped of its evolutionary implications.

By far Hubble's most remarkable bequest to modern science was his research showing that we live in an expanding universe. Hints of that phenomenon had begun to surface well before Hubble tackled the problem. During the 1910s, Slipher had noticed that most of the galaxies he observed seemed to be moving away from the sun. By 1924 a trio of astronomers—Carl W. Wirtz in Germany, Knut E. Lundmark in Sweden and Ludwik W. Silberstein in the U.S.—had assembled evidence that particularly small, and presumably distant, spiral nebulae are receding from the Milky Way faster than nearby ones. Their data were marginal, however, and Silberstein's polemical style and wild claims actually hindered the acceptance of the concept of an expanding universe.

Clearly, further progress required a bigger telescope; the 100-inch telescope on Mount Wilson was ideal for the task. In the mid-1920s Hubble was busy measuring the brightnesses of stars in spiral galaxies, so he enlisted his colleague Humason to determine the galaxies' radial velocities—that is, their motion toward or away from the observer. Humason calculated those motions by measuring the displacement of the emission and absorption lines in the galaxies' spectra. When an object is receding, its light is shifted toward the red end of the spectrum, resulting in a displacement called a redshift.

To search for signs of an expanding cosmos, Hubble needed to know not only the velocities at which galaxies are moving but also how faraway those galaxies are. Hubble's familiarity with measuring cosmic distances greatly assisted this effort. In each of the few galaxies whose distances he had determined by observing Cepheids, Hubble derived the mean absolute magnitude of the brightest stars and of the galaxy as a whole. He then extrapolated from these results to determine the distances to galaxies far beyond M31 and M33.

Hubble's first paper on a velocity-distance relation, published in 1929, sent shock waves through the astronomical community. The findings he presented became part of the basis for the big bang theory underpinning modern cosmology. In his paper Hubble emphasized that, in contrast to some previous researchers, he used "only those nebular [galactic] distances which are believed to be reliable." The paper combined his distance estimates with precision measurements of radial velocities collected by Slipher, Humason and others.

From these data, Hubble derived a linear relation between velocity and distance, written as $v = Hd$ in modern notation. The statement that the velocities of recession of distant galaxies are proportional to their distances is now known as the Hubble law, and the constant H is referred to as the Hubble constant. The Hubble law implies that the universe is expanding: velocities seem to increase as one looks progressively farther outward from any point within the universe.

Between 1931 and 1936 Hubble and Humason extended the law of redshifts to increasingly great distances. Hubble had immediately grasped that his findings supported the notion of an expanding universe, but he never completely accepted that the redshifts resulted only from the radial motions of galaxies. He consistently maintained that the motions inferred from the redshifts should be described as "apparent velocities."

Hubble estimated *H*—now understood as the expansion rate of the universe—to be about 500 kilometers per second per megaparsec (one megaparsec equals about three million light-years). In retrospect, Hubble's value of the constant was much too large. Astronomers today find values between 50 and 100 kilometers per second per megaparsec. Hubble's errors arose chiefly because he severely underestimated the absolute magnitudes of the brightest stars in the galaxies he observed. Nevertheless, it was a splendid first effort.

After his investigation of the law of redshifts, Hubble concentrated his efforts in observational cosmology. In particular, he sought to measure how the apparent density of extremely remote galaxies changes with distance, in order to determine the overall geometry of the universe. He collaborated with physicists Richard C. Tolman and Howard P. Robertson of the California Institute of Technology, who were at the forefront of research into theoretical models of the curvature of space. In this work, Hubble proved less than successful, mostly because of the difficulty in accurately measuring the brightness of dim, distant galaxies.

Although his later undertakings never equaled his remarkable findings made during the 1920s and early 1930s, Hubble continued to sway strongly the direction of astronomical research and to make that research better known to the public. His books *The Realm of the Nebulae* and *The Observational Approach to Cosmology* inspired a generation of young astronomers and physicists. Nobody who read Hubble's books or heard his lectures could ignore him; his compelling personality seemed less like those of other astronomers than like those of the movie stars and writers who became his friends in the later years of his life. Hubble often affected English clothes, expressions and mannerisms, habits that colleagues who knew of his Missouri background found unconvincing but that contributed to a dramatic public image.

In an effort to push the boundaries of observational cosmology, Hubble helped to organize the construction of a 200-inch Hale telescope on Mount Palomar in California. The advent of World War II delayed the construction of the telescope; soon after the Japanese attack on Pearl Harbor, Hubble joined the staff of the U.S. Army's Ballistics Research Laboratory at Aberdeen, Md., where he applied his early astronomical training to directing calculations of artillery-shell trajectories. The 200-inch telescope finally began operation in 1948. Hubble continued his research at Palomar until he died at the age of 63, felled by a stroke on September 28, 1953.

After Hubble's death, Humason and then Rudolph L. Minkowski used the 200-inch telescope to detect ever more remote, swiftly receding galaxies. Walter Baade recalibrated the period-luminosity relation for Cepheid variable stars and discovered that all Hubble's galactic distances, huge though they seemed, needed to be multiplied by a factor of two. Hubble's former student Allan Sandage, along with innumerable other researchers, has continued to refine the value of the Hubble constant. The triumphant work of Hubble's successors stands as eloquent tribute to the innovative thinking and boundless energy that enabled Hubble to redirect the course of modern cosmology.

SPIRAL GALAXY M33 was an early target of Hubble's efforts to determine the distance scale of the universe. Based on his observation of Cepheid variable stars in M33, Hubble proved that it lies far beyond the known limits of the Milky Way and hence must be an independent galaxy comparable to our own.

FURTHER READING

THE EXPANDING UNIVERSE: ASTRONOMY'S "GREAT DEBATE," 1900–1931. Robert Smith. Cambridge University Press, 1982.

MAN DISCOVERS THE GALAXIES. Richard Berendzen, Richard Hart and Daniel Seeley. Columbia University Press, 1984.

THE OBSERVATIONAL APPROACH TO COSMOLOGY: U.S. OBSERVATORIES PRE–WORLD WAR II. Donald E. Osterbrock in *Modern Cosmology in Retrospect.* Edited by R. Bertotti, R. Balbinot, S. Bergia and A. Messina. Cambridge University Press, 1990.

SELF-MADE COSMOLOGIST: THE EDUCATION OF EDWIN HUBBLE. Donald E. Osterbrock, Ronald S. Brashear and Joel A. Gwinn in *Evolution of the Universe of Galaxies: Edwin Hubble Centennial Symposium.* Edited by Richard G. Kron. Astronomical Society of the Pacific, 1990.

SUSTAINING THE AMAZON

by Marguerite Holloway, *staff writer*

Can scientists reconcile the inevitability of economic development with the preservation of rain forests?

Big, brown and sedate, two toads blink in the beam of the flashlight, while nearby pairs of green "eyes," the phosphorescent markings of click beetles, move lazily along the leaf-laden floor of the forest. On this night, Barbara L. Zimmerman is trying to find what she calls "the ultimate frog" for a group of visitors she is accompanying through the dark tangle of the Brazilian rain forest. The tropical biologist can hear her quarry's barklike call, but the vivid emerald creature, *Phyllomedusa bicolor,* is nowhere to be seen.

Instead the nocturnal search reveals the red irises of *P. tarsius,* the penultimate green tree frog; *Plica plica*, a mint-chocolate-chip-colored lizard; the two enormous, poisonous toads, *Bufo marinus;* as well as a tireless, resonant frog, *Osteocephalus taurinus,* that resides in the camp's drainpipe. The Amazon is like this: look for one thing, find another.

Zimmerman, who formerly stalked frogs in this forest 70 kilometers north of the city of Manaus to study their mating patterns, now pursues them primarily for pleasure. These days, her research centers on people, a shift in emphasis that is increasingly common in the Amazon. She has come here this time as part of a team of tropical scientists who are reviewing a project as well as possibilities for collaborative studies. Inevitably, their discussions turn to ways of reconciling the preservation of the rain forest with economic development.

Many of the scientists seem resigned to the prospect that the rain forest will, in large part, be cut down even as they struggle to understand the biological wealth of this tropical ecosystem. The five million square kilometers of

HARVESTING BRAZIL NUTS and other products may allow economic growth without destroying the Amazon. Researchers are debating the potential of such practices.

MARGUERITE HOLLOWAY spent a month in the Brazilian rain forest to report this article. Traveling from one end of the Amazon basin to the other, she interviewed scientists working at remote research stations, native preserves, extractive reserves, ranches and frontier logging camps.

Amazon tree cover make up the largest continuous expanse of tropical rain forest remaining in the world. Although such forests cover only 7 percent of the planet's land surface, they are inhabited by some 50 percent of the plants and animals found on the globe (estimates of which range from a total of two million to 30 million species). Most disturbing of all is the fact that nobody really knows how many. But as many as 27,000 species may be consigned to extinction every year, calculates Harvard University biologist Edward O. Wilson. This destruction is the result of forests being felled at a rate of 1,800 hectares (about 4,500 acres) every hour.

The impact of these extinctions is multifold: the rain forests have profound philosophical, spiritual, cultural, scientific and economic significance. Because Brazil is considered to have the highest species diversity on the earth, it is the epicenter of efforts to stem deforestation. So far the country has lost only an estimated 12 percent of its forest, a sign of hope to many researchers and development experts. Yet the country has a debt burden of $121 billion, staggering poverty and one of the most unequal distributions of wealth in the world. Any successful attempt to preserve the forest must necessarily meet human economic objectives.

By working with the people who live in and use the forest—including Indians, rubber tappers, ranchers, loggers, farmers and industrialists—scientists such as Zimmerman hope to strike a balance between extremes of conservation and exploitation that will protect species and threatened environments while fostering economic development and reducing poverty. "Do what you can do—just do something," urges Don E. Wilson, head of the biodiversity program at the Smithsonian Institution.

These efforts are among the myriad attempts to implement what is called sustainable development. The idea was the linchpin of the 1992 United Nations Conference on Environment and Development, but the fashionable phrase, which now appears in nearly every grant or loan application, is still being defined. Thus, researchers are seeking pieces of a mosaic of development, the overall pattern of which, like the web of biodiversity, can only be guessed at.

Slowly, however, some scientific underpinnings for sustainable development appear to be emerging. Sometimes the answers seem to contravene assumptions about what might be ecologically sustainable. Scientists are investigating how to use the millions of hectares of formerly productive pasture that lie abandoned throughout the Amazon. Once cut down, the rain forest does not necessarily become a desert but rather a field or a secondary forest that can again be harnessed for ranching, logging or agriculture. If this land can be returned to productivity, it would reduce pressure to cut virgin forest. Indeed, a group of researchers recently challenged dogma by suggesting that cattle ranching—long thought to be an ecological and economic anathema—may be a viable land use.

Whatever the project or finding, controversy is never far off in Brazil. The country is so vast and the economic, social and political problems so complicated that every issue is labyrinthine: follow the thread, and there are often several Minotaurs. One of the most contentious topics involves the future of indigenous people and their land. It is on one of these Indian reserves, with a people called the Kayapó, that Zimmerman has chosen to work.

Zimmerman and her sponsors—the environmental group Conservation International and the David Suzuki Foundation, a Canadian nonprofit organization—are setting up a scientific research station on remote Kayapó territory in the southern reaches of the Amazon. The villagers of Aukre, which means "place where the river bottom is noisy," have set aside 5,000 hectares for the field station within the 500,000 hectares of rain forest that they control.

The project will characterize the biodiversity of a region that has not been extensively studied, Zimmerman explains. The flora and fauna of the Amazon vary so greatly that adjacent sites can shelter entirely different species. One hectare in the eastern Amazon may have 120 tree species, whereas a nearby hectare may have 170. The woods around Aukre may have creatures or plants that are not found, or are infrequently seen, elsewhere in Brazil.

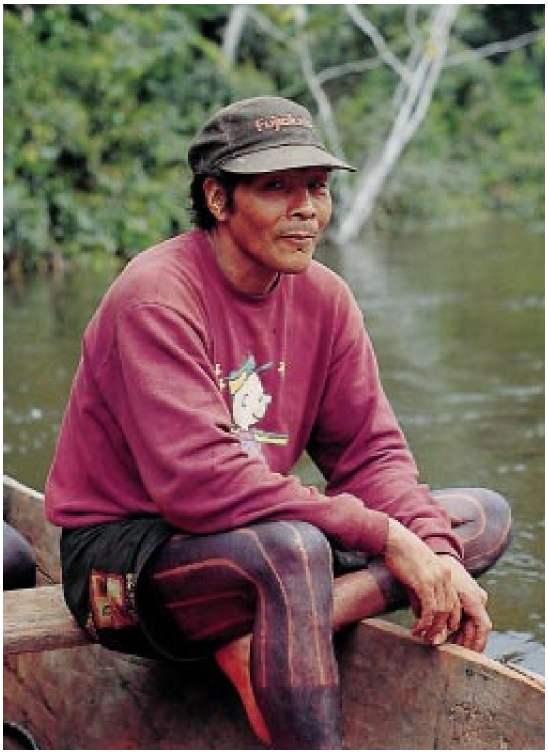
Many scientists have shied away from working with Indians because the issues

AMAZON RAIN FOREST extends into nine countries and covers 60 percent of Brazil—some 510 million hectares (about 1.3 billion acres).

facing them are so volatile. “It is a land mine,” comments Anthony B. Anderson, program officer at the Ford Foundation in Brazil. “There are politics within the tribes and without, as well as all of these misconceptions about what the Indians are supposed to be.”

The debate about what indigenous people should do with their land is bitter. As many as 79 million hectares have been or are in the process of being designated as reserves, according to Stephan Schwartzman of the Environmental Defense Fund. To many scientists and environmentalists, this expanse represents a perfect series of places to preserve and study biodiversity as well as native cultures. The Kayapó, for instance, control 10 million hectares, a region approximately the size of Ontario.

Yet these lands are also rich in minerals and timber and so are ripe for economic development. Many are illegally infiltrated by loggers and gold miners, such as the territory of the Yanomami in northern Brazil, where violence, disease and mercury poisoning, the last a result of the gold extraction process, are pervasive. But the fierce reputation of Kayapó warriors has successfully deterred encroachments. Some Kayapó chiefs have opted to sell mahogany and mining rights, a decision that has aggravated environmentalists. Greenpeace, the international environment group, recently sought to enforce a ban on mahogany logging on the tribal lands—an action that has, in turn, aggravated certain Kayapó. It is estimated that about 40 percent of the Amazon's mahogany trees—which, in general, only grow in a strip across the lower portion of the rain forest—thrive on reserves.



KAYAPÓ TRIBESMAN, Bastian, waits in a canoe that transports him between a forest reserve and his village, Aukre. He is working on a field station that may bring researchers and revenue to Aukre.

Noble Savage

Development experts and conservationists, many of whom struggle to help establish reserves, expect that Indians should maintain the rain forest intact. This view has been shaped, in part, by anthropological studies documenting tribes' sustainable use of the forest and their sophisticated agricultural systems. Others experts argue, however, that sustainability has been achieved by default: the forest is large enough to recover from alteration, and populations are small enough to have limited impact. “The idea of the noble savage has really confused the issue,” Anderson notes. “Indians, like everyone else, are opportunistic. In the past, they used resources sustainably, but not always.”

If Zimmerman's project is successful, it could help settle the debate by supplying a source of income for Indians while simultaneously maintaining biodiversity. Aukre is an interesting choice for such a site because it has experienced each facet of the controversy. The village's chief, Paiakan, is famous as a spokesperson for protecting rain forests and tradition, as well as for being a skillful politician who could unify previously warring tribes in the face of developers. Yet other villagers have disagreed with his view. In addition, Paiakan can no longer easily be an advocate for indigenous people in the Amazon: he has reportedly been charged with attempted rape and is awaiting trial.

The region around the reserve is as scarred as the tribe. The flight northwest from Redenção—a frontier town whose economy is based on minerals and mahogany from Kayapó land—to Aukre offers a view of the entire spectrum of development. Ranches, logged plots and rivers disfigured by gold miners are slowly replaced by uninterrupted forest. On the ground, the new research site appears intact, although it is bordered on one side by an old logging road. The station stands in the middle of a clearing just off a river, an hour by canoe from Aukre. Bright-blue morpho butterflies flit down trails pockmarked with tapir tracks, and the wolf whistle of screaming pia birds rings out from the canopy.

Zimmerman intends to fill this place with graduate students and researchers who will study the forest and work with the Kayapó. She notes that natives have a different way of knowing the forest—for instance, men called *mateiros* can identify hundreds of trees by smell alone. She expects that such collaboration will yield medicinal plants or new crops and perhaps a strategy for the sustainable harvesting of mahogany.

In turn, the project is expected to bring income to Aukre. Scientists will buy local products, ecological tours may visit the site under Zimmerman's stewardship and fly-fishermen may explore the river's varied population of fish. Collaborative studies with the Kayapó could also lead researchers to medicinal plants or agricultural techniques that bring some financial return. Although the town makes money harvesting Brazil nut oil for Body Shop International, a British company that develops and sells natural beauty products, it is not enough.

The fervor about the future of Indian lands is echoed in arguments about another kind of reserve: extractive reserves. These areas—which are also set aside by the government and whose boundaries are also rarely enforced—are places where local people harvest natural products, including latex, palm hearts and Brazil nuts, without damaging the rain forest. Such reserves were made famous by Francisco Alves Mendes Filho, known as Chico Mendes, a rubber tapper from the town of Xapuri in the western state of Acre. Because of his efforts to organize the rubber tappers to fight for their land, Mendes was murdered by cattle ranchers in 1988.

Extractive reserves, of which there are now 14 covering a total of three million hectares, confer land rights to squatters and people whom large landholders have dispossessed. (In Brazil, 4.5 percent of the landowners control 81 percent of the farmland.) The presence of these powerful ranchers is felt for the entire four-hour drive from the city of Rio Branco to Xapuri. The road is flanked by pastures, and the landscape is verdant but eerie: ghostlike Brazil nut trees, lone relics of the forest, stand in the green fields, their bleached branches luminous against the backdrop of dark, stormy sky. A law prohibits cutting these majestic trees, but once pasture is burnt around them, the Amazon sun inevitably desiccates them.

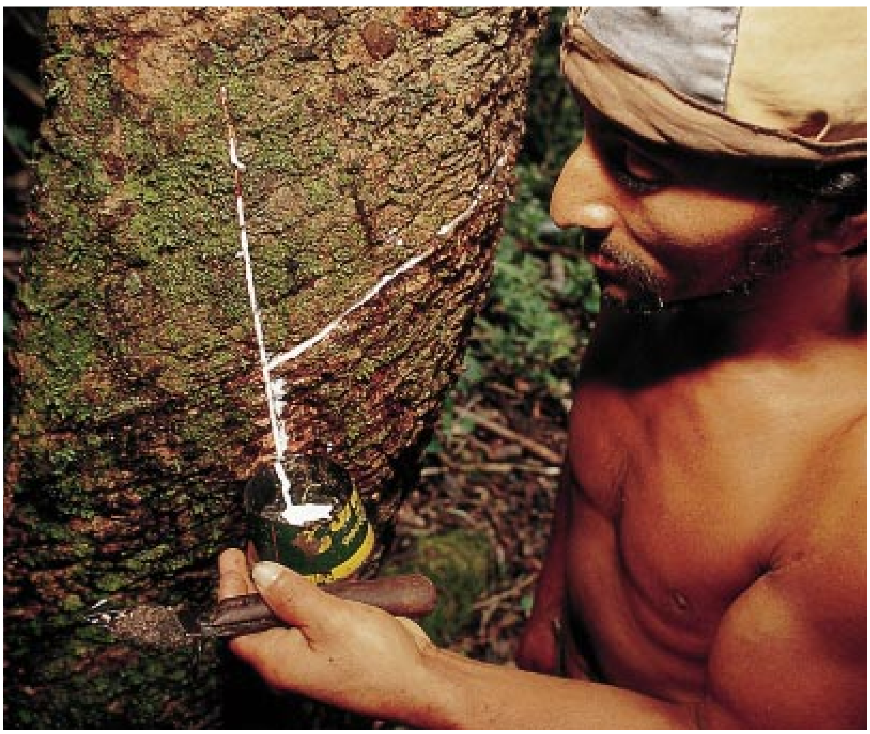
Disagreements about extractive reserves pivot on questions of economic viability. Scientists on one side argue that insufficient markets exist for many of the products. By studying past market cycles, Alfredo Kingo Oyama Homma of the Brazilian Enterprise for Agricultural Research (EMBRAPA) found that when a forest product such as latex becomes commercially important it is inevitably introduced into higher-yielding plantations; in 1991 about 60 percent of Brazil's natural rubber came from plantations. Or the material is made synthetically. As a result, the price plummets, and small-scale extraction ceases to be profitable.

Other researchers counter that the matter has not been settled. One study conducted in Peru several years ago found that 72 products could be harvested from a one-hectare plot for an annual yield of $422. The investigators concluded that this approach was more lucrative than the one-time logging profit of $1,000. On the other hand, Anderson notes, "if you took a random sample of forests here in the Amazon, you would not find many that would have such a high value of nontimber forest products." In addition, the Peruvian products could be marketed regionally, a situation that does not yet exist for many of the reserves in the Amazon.

Nevertheless, Anderson and others argue that there may be a compromise. "I am one of the proponents, so I am suspect," Anderson cautions. "But I like to think that I am somewhere in the middle ground. Maybe the term 'extractive reserve' is a misnomer. These are areas zoned to protect the rights of people who live in the forest." Anderson explains that once land claims are guaranteed, extractivists begin to look for supplemental livelihoods—such as raising livestock or planting additional crops. Extraction "is undergoing a transformation, and therefore it is potentially a basis for sustainable development." The implications of this economic diversification for the preservation of biodiversity are not yet established.

Many scientists and nongovernmental organizations—Cultural Survival, a Boston-based group of anthropologists, and the New York Botanical Garden's Institute for Economic Botany, among them—are working with rubber tappers and Brazil nut gatherers to find other products, such as medicinal plants and fruits, and to open new markets. Some are addressing the overwhelming problems of communication and transportation. Many individuals who harvest forest products travel for days to sell them.

Getting out of Xapuri, even at the outset of the rainy season, illustrates the challenge of reaching the market. Gomercindo Clovia Garcia Rodrigues, a member of the extractivist cooperative of Xapuri, hitches a ride with a group of travelers back to Rio Branco so he can visit his dentist—he says he was shot in the jaw by ranchers a while ago, and the wound continues to hurt. As he describes the woes of driving at this time of year, the white Fiat sinks in the red muck of the road, for the first time. After the fourth escape, we are covered with a thick, rust-colored patina of mud. The windshield wiper breaks, the rain blinds and the battery fails. Rodrigues laughs: the Amazon has made his point.



RUBBER TAPPER extracts latex. As the pattern of diagonal lines shows, tappers repeatedly cut each of the rubber trees on their land to collect the white liquid.

Amazon Microcosm

At the opposite end of the country, in the eastern state of Pará, researchers are working with Rodrigues's traditional foes: ranchers and loggers. In a valley capped by smoke, fires from dozens of sawmills smolder. The black and gray plumes hang in the humid air of a late afternoon. The only motion is that of children playing alongside the barren road, tin cans tied to the bottoms of their feet like shiny platform shoes.

Although the image of a deforested Amazon is by now a familiar one, a close look at this landscape reveals an unusual collaboration. Christopher Uhl of Pennsylvania State University, Daniel C. Nepstad of the Woods Hole Research Center and Adalberto Verissimo, Paulo Barreto and their colleagues at the Amazon Institute of Man and the Environment (IMAZON) are working with owners and employees at local sawmills and farms. They are documenting ways of reclaiming land that has been abandoned and of managing forests for timber. Uhl and Verissimo are fond of describing the region near the town of Paragominas, where they have focused their efforts, as a microcosm for the Amazon. They believe that if the forces driving deforestation are understood and altered here, the lessons can be applied to states such as Rondônia and Mato Grosso, which, by 1988, had both lost some 24 percent of their forest.

Pará encompasses one third of the Brazilian Amazon and had lost 13 percent of its forest cover by 1990. The area opened for development in the mid-1960s, when completion of a highway connected the city of Brasília to Belém, the hub of Brazil's rubber boom. At that time, the government promoted cattle herding by offering subsidies—often as great as 75 percent of the costs of starting the ranch—to settlers. The government also offered property to anyone who could demonstrate that they had cleared land, a policy that led to spasms of deforestation.

Several problems quickly emerged.

BOLUSES OF LATEX are stacked in a courtyard at the marketing cooperative in the town of Xapuri after they are brought in from the forest.

Because the forest covets its nutrients, its soils are relatively poor. Although ranchers slashed and burned the forest, releasing phosphorus, nitrogen, potassium and other riches into the ground, the infusion was short-lived. After a few seasons of overgrazing, grasses no longer grew. A riot of unruly and, to cattle, unappetizing vegetation soon covered the fields. By the 1980s ranching was not economically worthwhile. The government proved as unsupportive as the soil. A fiscal crisis dried up subsidies, making clear-cutting and cattle raising even less attractive.

A new industry emerged in ranching's stead. The native forests of southern Brazil had by then been for the most part leveled, and ranchers realized they could afford to rejuvenate degraded pastures by selling logging rights to their property. More than 100 species of wood, which had been viewed as impediments to pasture, were harvested. By 1990 there were 238 sawmills near Paragominas.

Today logging is the area's principal industry, and it is just as unsustainable as ranching was. IMAZON researchers have determined that for each tree that is removed in a process called selective logging, 27 other trees that are 10 centimeters or more in diameter are severely injured, 40 meters of road are created and 600 square meters of canopy are opened. Recovery of the forest is slow even under these conditions: it can take 70 years for selectively logged forests to resemble their original state.

Nevertheless, Verissimo, Uhl and others think sustainable approaches to logging and ranching are possible. "I am much more optimistic than I was five years ago," Verissimo comments as he bounces in a car on the way to a ranch, Fazenda Vitória. At the study station there, he climbs a tower that stands at the intersection of the four distinct kinds of land that characterize the region: primary, or pristine, forest; partially regrown, or secondary, forest; pasture that is in use; and abandoned pasture. Verissimo points to the old pasture, explaining that the forest is slow to invade such sites, in part, because some seed-dispersing birds do not venture into open spaces.

Although there are about 10 million hectares of abandoned pasture in the Brazilian Amazon, scientists are just beginning to monitor what happens to land once it has been burned, grazed and then left. The investigations support one of the most exciting possibilities for sustainable development. If already deforested areas could be efficiently used for agriculture, ranching or logging, pressure could be taken off the primary rain forest. "There is no need to deforest when two thirds of the already deforested land is not being used in a reasonable way," Uhl notes.

Alternatively, if such lands could be restored or coaxed back into forest, biodiversity could be better protected. Understanding which trees, plants and creatures reestablish themselves first and under what conditions could allow researchers to speed up the recovery. Uhl cautions, however, that making the process economically attractive is critical. "Restoration makes us all feel good, but it doesn't necessarily address the key issue," he warns. "Putting money into it without economic development is not the best thing to do."

Studies of pastures and ranching have already led Uhl and his colleague Marli M. Mattos of Woods Hole to the somewhat heretical conclusion that, under certain conditions, cattle ranching can be profitable. They observe that if ranchers manage abandoned land intensively—that is, by introducing fertilizer or by planting better adapted species—cattle herds for beef and for dairy products can be lucrative. In fact, ranchers could increase their profits as much as fivefold at the same time that employment could rise. (Given estimates of

MEASURING PRIZE HARDWOODS is one of the first things a logger does to select the tree he will cut, demonstrates Paulo Barreto (*right*) of the Amazon Institute of Man and the Environment (IMAZON).

population in Brazil—246 million by 2025—considerations of employment are fundamental.)

The two scientists note that the drawbacks to intensive ranching are significant. "Fire is becoming more and more common," Uhl states. "It will increase as landscapes are altered. The high forest is by nature quite resilient to fire, but as soon as you begin to alter it, it is much more prone to burning." According to Nepstad of Woods Hole, more than one half of the land in an 80-kilometer radius around Paragominas burned in 1992.

There are other ecological costs as well. Nepstad has shown in studies of soil and root systems that pastures hold only 3 percent of the aboveground carbon and 20 percent of the root carbon that forests do. Moreover, some 220 tons of carbon are released when the slash from each hectare of forest is burned. This emission has implications for global warming: tropical forest loss is estimated to have contributed one quarter of the carbon released into the atmosphere during the past decade.

Evapotranspiration—the means by which roots absorb water from the soil and release it—is also reduced in pasture because the root systems are relatively shallow and cannot reach deep deposits of moisture. In the Amazon, where 50 percent of precipitation may be generated by the flora, changes in evapotranspiration could cause alterations in local climate and rainfall patterns that adversely affect agriculture. Reductions in rainfall, in turn, cause the local temperature to increase, hampering vegetation and crops.

The ranching conclusions reveal some of the problems of defining the word "sustainable." "It is a really tricky term. There are all sorts of ways to measure sustainability, from an environmental, an ecological and a social perspective. It is a rare land use indeed that could fulfill all three criteria," observes Anderson of the Ford Foundation. "Don't get me wrong, I am not saying ranching is good, but it is not necessarily as bad as it seemed in the earlier literature. The question is not whether or not ranching is suitable, because it is there to stay."

Research on sustainable logging in Paragominas is not as controversial. In a forested site adjacent to the study tower, IMAZON's Barreto describes one simple strategy he has designed to help combine forest management and profit. Loggers routinely send workers into the forest to mark and cut the prize trees. Equipment then follows and creates roads. Barreto has found that by plotting the chosen trees in a simple software program, a shorter—and therefore less destructive and less expensive—network of roads can be designed.

In addition, injury to the surrounding trees can be reduced by as much as 30 percent. Pruning vines in a selectively logged forest can also permit light to reach seedlings, accelerating the growth of commercially important species. If loggers implemented these two techniques, Barreto says, forests would take 35 instead of 70 years to reach their prelogged stature. "It would reduce their profits only slightly to get sustainability," Verissimo asserts.

Because internal and external demand for Amazonian timber is expected to grow, more research on forest management strategies is needed. In 1976, 14 percent of the country's sawn wood came from the rain forest; by 1986 the share had grown to 44 percent. The number of wood processors grew from about 100 in 1952 to more than 3,000 in 1992. Unlike domestic demand, however, international demand may catalyze changes in logging practices. Currently less than 5 percent of the world's supply of tropical hardwood comes from Brazil, but supplies in Asia are plummeting. Yet "outside demand for timber could be quite positive if nongovernmental organizations are involved in promoting green labeling," Uhl says. "Retailers want to know where this wood is coming from. This little step is tremendously important."

Regardless of findings such as those in Paragominas, research amounts to nothing if it is not implemented. Because Brazil has no coherent Amazonian policy, the IMAZON workers and their colleagues have suggested that countywide plans would be effective. The Brazilian Constitution of 1988 conferred more power to counties; therefore, these local governments could enact legislation requiring that specified land be used intensively for pasture and logging. Such laws would generate jobs for the region and preserve primary forest. Uhl says ranchers and mill owners are beginning to implement aspects of this management strategy as they realize that resources are finite and that well-placed investments in intensification will help guarantee sustainability.

Survival, rather than profit, is the consideration behind another principal cause of deforestation. For impoverished farmers, shifting cultivation often provides the only source of livelihood, as the ubiquitous small, smoldering plots of trees alongside roads testify. "Natural resources are the bleeding valve for the poor; that is where they turn to survive," observes Ralph M. Coolman, a tropical ecologist at North Carolina State University, who is currently working with EMBRAPA.

On the fields of an old water-buffalo ranch and former rubber and palm plantation near Manaus, Coolman and his colleagues are investigating agroforestry systems. Agroforestry combines animal husbandry with farming and forestry. The practice has a long history, but "scientific agroforestry is relatively new," says Erick C. M. Fernandes, also of North Carolina State and EMBRAPA. By determining which plants are best suited to particular soils, the researchers hope to be able to increase the subsistence and profit yield of shifting cultivars, while letting them stay indefinitely on the deforested land. They are identifying species that are best suited to fixing nitrogen and replacing other nutrients.

The effort to balance preservation and economics that characterizes the work of Fernandes and Coolman, the Paragominas group and Zimmerman's team can also be seen at one of the longest-running programs in the rain forest. This study, the brainchild of biologist Thomas E. Lovejoy, was created in 1979 as a collaboration between the National Institute for Amazon Research (INPA) and the World Wildlife Fund. Originally called the Minimum Critical Size of Ecosystems Project, it is now administered jointly by INPA and the Smithsonian Institution, where Lovejoy serves as assistant secretary for external affairs.

Business and Biodiversity

The program was established to answer a fundamental question of ecology: What is the smallest size an island (or, in this case, a tract of forest) can be and still maintain its biodiversity? The findings have direct implications for protecting species. Because landowners in the Amazon are required by law to retain 50 percent of their forests, scattered woods of all different sizes and shapes can be found everywhere, often surrounded by pasture.

By studying these various-sized plots and monitoring the changes in species composition, Lovejoy and his colleagues substantiated the prediction that larger

SAWMILL NEAR PARAGOMINAS, a town in the eastern Amazon, is one of 238 in the vicinity. Although Brazil supplies less than 5 percent of the world's hardwood, this share is expected to increase as stocks in Asia dwindle. Researchers are working with loggers and sawmill owners here to determine how to practice sustainable logging.

forest fragments harbor more species than do smaller ones. But they also found that size alone was not always the determining factor. Instead it was crucial to understand the habitat requirements of each species. The project provided evidence that corridors—wooded paths that connect one fragment to another—facilitated the movement of certain species between fragments. Because it is hard to find large, uninterrupted parcels of land in many areas throughout the world, corridors are now incorporated in plans for wildlife preserves.

The project recently included an additional focus: the study of restoration, including the regrowth of pasture. Its new name, the Biological Dynamics of Forest Fragments Project, reflects this emphasis. Graduate students and researchers are expected to supply data for forest management, soil reclamation and commercial harvesting. The shift also ensured continued funding. According to tropical scientists, money for sustainable development is easier to obtain than are grants for basic research.

Lovejoy believes that another key to saving the rain forest lies in working with business. In the mid-1980s he developed the idea of debt-for-nature swaps. Recognizing that nations burdened with foreign debt could devote few resources to protecting ecology, Lovejoy proposed that conservation organizations repay a portion of the debt in exchange for a guarantee that land would be protected. Alleviating debt would both maintain biodiversity and strengthen the economy.

Building on this idea, Lovejoy and others are drawing attention to the financial importance of biodiversity. For example, pharmaceutical products, genetic stocks for agricultural products and timber species can be found in these forests. By publicizing this economic potential and the importance of conserving the resource base for future profit, Lovejoy says companies can be brought into the environmental loop—an approach that, he admits, is perceived with wariness on the part of many "green" organizations. Lovejoy, who in the course of one day easily moves from sitting on the dirt floor of a hut in a Kayapó village swatting blackflies to dining with the state's governor at a table dressed with crystal and silver, argues that sustainable development and the protection of biodiversity must be practiced at all levels of society.

To this end, he is advising the Brazilian Foundation for Sustainable Development, a consortium of 24 of the largest Brazilian companies, formed after the 1992 Earth Summit in Rio de Janeiro. The foundation plans to support research that will define sustainable development and to help companies act in environmentally sound ways. It will also finance the Forest Fragments Project. "Although the first step for sustainable development is in local communities, it is not only for them," says Angelo dos Santos, a foundation ecologist who has worked with INPA and Lovejoy. "Because of global climatic change, we need to work with big enterprises. We need a whole new approach."

At a December meeting at Camp 41—the "Tiffany" of sites at the Forest Fragment Project because of its ample hammocks and good cooks—Lovejoy, Zimmerman, Wilson and biologists from Brazil, Panama, Costa Rica, Peru and Mexico met to evaluate the project and to consider future studies. David B. Clark of La Selva Biological Station in Costa Rica pointed out that the only large tracts of forest left in his country are in parks. "I am trying to be philosophical about it," Clark says. "Most reserves in South America are going to be in this situation. Development is going to be right up against the protected areas."

As the scientists walked through the rain forest, noting differences and similarities between this region and those they study, it was apparent that no one was able any longer to devote attention exclusively to the study of biodiversity. "Many of us were not trained as administrators but as research biologists," the Smithsonian's Wilson comments. "Yet we find ourselves, out of necessity, doing conservation work." If they do it well enough, human beings and the rest of Amazonian nature may yet find a way to survive together. After all, in the Amazon you look for one thing and keep finding others.

BIOLOGIST THOMAS E. LOVEJOY stands in the rain forest that he first visited in 1965. His Amazon research and, more recently, his advisory work have increasingly sought to find economically successful ways of preserving biodiversity.

FURTHER READING

THE FATE OF THE FOREST: DEVELOPERS, DESTROYERS AND DEFENDERS OF THE AMAZON. Susanna Hecht and Alexander Cockburn. HarperPerennial, 1990.

BIODIVERSITY. Edited by E. O. Wilson. National Academy Press, 1992.

THE LAST NEW WORLD: THE CONQUEST OF THE AMAZON FRONTIER. Mac Margolis. W. W. Norton and Company, 1992.

SUSTAINABLE AGRICULTURE AND THE ENVIRONMENT IN THE HUMID TROPICS. National Research Council. National Academy Press, 1993.

Try, Try Again

Making antibodies more useful by making them more human

"Cloned antibodies promise medical revolution," screamed the front-page headline of the *New York Times* science section on August 5, 1980. Throughout the next decade, dozens of biotechnology firms struggled to turn monoclonal antibodies (MAbs) harvested from mouse tumors into new drugs to treat everything from cancer to bacterial infections. They have yet to deliver. Despite years of intense research, only one MAb-based therapeutic drug, OKT3 from Ortho Pharmaceuticals, has gained Food and Drug Administration approval for the U.S. market.

But enthusiasm for monoclonals seems to be enjoying something of a renaissance. A handful of companies are now perfecting techniques for manufacturing fully human and "humanized" antibodies that promise to surmount some of the problems that have plagued murine monoclonals. They are buoyed by results from early clinical trials of this new generation of antibodies against a variety of diseases.

The major obstacle to medicinal use of murine antibodies proved to be the very immune systems they were intended to bolster. Many patients produce human anti–mouse antibodies (HAMA) when they are given shots of murine monoclonals. The ensuing battle between the mouse and human proteins reduces the drug's effect, quickly clearing it from the bloodstream and often causing allergic reactions. Another problem, explains Robert A. Fildes, chief executive officer of Scotgen Biopharmaceuticals, is that "when you put a mouse antibody into a human being, it does not turn on the human immune system as effectively as a human antibody would."

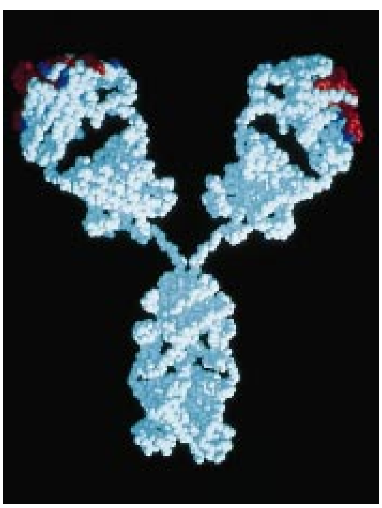
Human antibodies have proved difficult to make outside the body, however. Mouse monoclonals, the Nobel Prize–winning discovery of César Milstein and Georges Köhler in 1975, are made by fusing a mouse's *B* lymphocyte—a white blood cell in the spleen that produces antibodies but is difficult to culture—with a cancerous melanoma cell to produce an immortal hybrid cell, or hybridoma. Such cells create antibodies and replicate furiously in a fermentation tank. Murine antibodies can thus be mass-produced for nearly any antigen injected into a mouse. Unfortunately, however, human *B* cells are loath to fuse with cancerous cells.

The ability of murine MAbs to home in on a particular cellular target won them a niche in the lucrative market for diagnostic tests and cancer imaging. Monoclonals that bind to proteins on the surface of cancer cells can be attached to radioisotopes to help doctors see clusters of malignant cells invisible to x-rays, CT scans and magnetic resonance imaging. And because imaging agents need be given just once or twice, HAMA reactions are less of a problem. Even so, only one imaging agent, a murine monoclonal for diagnosing colon cancer, has been approved to date. "Look back over 10 years: there has been huge success with mouse MAbs as in vitro diagnostic tests—it's a billion-dollar industry worldwide," Fildes says. "But in that same period, virtually every company that has tried to develop mouse antibodies for therapy has failed miserably."

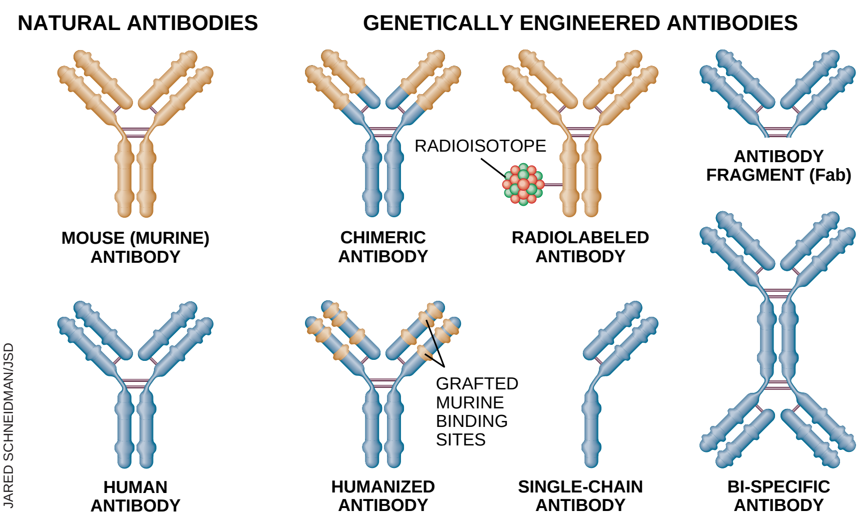
Witness Xoma, which was within shouting distance of winning FDA approval for its MAb-based sepsis drug last year before final clinical trials failed

Antibody Anatomy

Antibodies are variations on a theme. Natural antibodies (*left*) are proteins composed of amino acid chains twisted into complex three-dimensional shapes. They lock onto antigens—target molecules—that match the shape and chemistry of the binding sites (*shown in red*) on their arms. Geneticists can convert mouse antibodies to various "humanized" forms (*right*) by combining the genes for murine binding sites with those for a human framework.



PROTEIN DESIGN LABS, INC.



JARED SCHNEIDMAN/JSD

Monoclonal Antibody Therapeutics in Development

Murine Antibodies

IDEC Pharmaceuticals: Phase III trials* for *B* cell lymphoma.

ImmunoGen: Phase III for various cancers; phase I and II for small-cell lung cancer.

Immunomedics: Phase I and II for colorectal cancer and *B* cell lymphoma.

NeoRx: Phase I for lung and ovarian cancer.

Humanized Antibodies

Cytogen: Preclinical evaluation for breast, lung and ovarian cancer.

Genentech with NeoRx: Phase II for breast and ovarian cancer.

IDEC Pharmaceuticals: Preclinical testing of "primatized" macaque/human antibody for rheumatoid arthritis.

Protein Design Labs: Phase II for graft versus host disease; phase I and II for acute myeloid leukemia; preclinical tests for lung, colon and breast cancer, inflammation, lupus and multiple sclerosis.

Scotgen Biopharmaceuticals: Preclinical testing of 12 antibodies against cancer, cytomegalovirus, HIV and atherosclerotic plaque.

Human Antibodies

Protein Design Labs: Completed phase I and II for retinitis and hepatitis B; preclinical tests for herpes and chicken pox.

*Marketing approval of drugs by the Food and Drug Administration generally requires three phases of human clinical tests: the first for safety and toxicity, the second for efficacy on a small scale and the final for large-scale, statistically significant effectiveness.

to show sufficient effectiveness. Xoma has since lost its CEO, shut down clinical trials of four other products and slashed its staff by 25 percent. Nevertheless, the size of the pot of gold at the end of this rainbow has induced many companies to continue working on therapeutic MAbs; industry analysts peg the potential market at $3 billion for cancer alone.

Most biotech firms now seem to agree that therapeutic antibodies will have to beat HAMA to succeed. Here is where it helps most to be human. Actually, an antibody need not be entirely human to fool a patient's immune system. Simply replacing the constant region, or stem, of the Y-shaped murine antibody with its human equivalent produces a chimeric antibody that may sneak through undetected. Greg Winter of the British Medical Research Council improved on that idea in the mid-1980s by splicing the genes for just the binding sites of a mouse antibody alongside the genes for the rest of a human antibody to make a "humanized" antibody.

Getting the humanized version to lock onto the same antigens as the original takes some work, however. "You sometimes have to tinker with the amino acid sequences in the framework of the humanized antibody in order to get the binding sites in the right three-dimensional form," says Fildes, whose company has used Winter's technique to humanize more than 20 antibodies taken from mice. "The trick is to know which ones to change—it's really an art."

A start-up company in California is making a science of that art. Protein Design Labs (PDL), which went public early in 1992, has used protein-modeling software to compare the amino acid sequence of the binding sites with a library of other antibodies whose spatial structure has been determined by x-ray crystallography. Once the model picks out which mouse amino acids to keep and which to replace, PDL synthesizes the new genes, inserts them into mamalian cells and cultures the cells to produce the humanized antibodies.

So far PDL has humanized more than a dozen antibodies against viruses, cancers and even adhesion molecules that play a role in inflammation. In the three clinical trials PDL has conducted to date with its humanized MAbs, "we have not seen any HAMA response," claims Peter Dworkin, PDL's spokesperson.

The fastest arrows in PDL's quiver are not its 90 percent human monoclonals, however, but entirely human MAbs produced by cell lines that the company purchased from Sandoz in April. From the spleen of a car crash victim, Sandoz managed to isolate human *B* cells that produce antibodies against cytomegalovirus, the cause of an eye infection in 20 percent of AIDS patients that often leads to blindness. By fusing this *B* cell with a human/mouse hybridoma, Sandoz created "triomas," which thrive in culture and produce many different antibodies against the virus. For $4.5 million, PDL got four such triomas, two of which—anti-cytomegalovirus and anti–hepatitis B—have already completed initial clinical trials for toxicity and efficacy.

Triomas still depend too much on luck and labor to be widely useful. GenPharm International thinks it has a better way to produce human MAbs. Using its experience producing transgenic animal models for cancer and cystic fibrosis research, the company engineered a mouse that produces only human antibodies. Researchers turned off genes for each of the mouse's two native antibody chains and then inserted and activated genes for human antibody chains. The result is a mouse strain that "is able to make any human antibody, including antibodies to human antigens," says Jonathan MacQuitty, the company's CEO.

GenPharm also recently announced that after four years and $10 million it had created a second line of human MAb-producing mice by using artificial chromosomes genetically engineered in yeast cells to transfer huge blocks of DNA from humans to the rodents. Unlike the rest of GenPharm's products, "we won't sell these mice," says Howard B. Rosen, director of corporate development. "We will work with researchers on a collaborative basis. It's sort of the mouse that lays a golden egg—we'd hate for people to just run off with it."

Researchers who have isolated a human antibody may soon use a technique refined at the Scripps Research Institute to mass-produce it in bacteria. Scripps molecular immunologist Dennis R. Burton amplifies the antibody's DNA and clones it into "display" viruses. "Each virus is a little unit, with one antibody gene inside and the protein coded by that gene on the outside," Burton explains. When poured onto a plate coated with the target antigen, viruses containing antibodies that match that antigen stick. The rest are washed away. Burton then infects *Escherichia coli* with the selected viruses, converting them into bacterial antibody factories. The process does have a serious limitation, Burton concedes. Because bacteria lack the sugars needed to make the stem of the antibody, they reproduce only the arms—a fragment known as Fab. "Although this fragment alone will neutralize a lot of viruses, we don't know how it works," he admits.

Even so, several biotech firms are looking at antibody fragments with new-found respect. A whole antibody is too large to penetrate the surface of many tumors; fragments may be able to go places whole MAbs cannot. But fragments are removed from the bloodstream in hours rather than days. "It's a two-edged sword," says Thomas J. McKearn, president of Cytogen. "An intact MAb will pass through the circulatory bed of a tumor many more times than will a fragmented version."

Researchers at Scotgen and elsewhere are also joining two different antibodies to make X-shaped "bi-specific" antibodies that attach to a diseased cell on one end and an immune cell on the other, dragging immunologic soldiers to the enemy. Medarex recently began clinical trials on such a drug for breast and ovarian cancer. "The problem is making these things," Fildes complains. "It's a long, tortuous and very expensive process."

Indeed, all these new technologies are considerably more expensive than churning out old-fashioned murine MAbs. And there are no guarantees. Centocor's human antisepsis monoclonal fared little better than Xoma's murine version in final efficacy trials. After investing $250 million in development, Centocor has seen just $31 million in sales. McKearn, for one, takes this as a sign that it is too early to dismiss mouse monoclonals. "The literature is replete with examples where people have ignored common sense and simply jumped on the bandwagon of humanization," he says. But of course sometimes the bandwagon goes straight to the party. —*W. Wayt Gibbs*

Earcons

"Audification" may add a new dimension to computers

In the early days of computers, an enterprising programmer would often place a portable radio on top of one of these primitive machines. The internal processing circuitry would act as a transmitter whose signals could be picked up a block away if a radio were tuned to the proper frequency. Each program would send out a characteristic stream of whistles and clicks that indicated that the software was running as it should.

These days, history is repeating itself. Software engineers struggling to program parallel processors have enlisted the help of electronic sound synthesizers to decipher the effectiveness of their improvisations. A bell tone or a violin note may reveal that processors 13 and 15 are not running, letting the programmer then rework the software to keep them busy.

Computer scientists, in league with electronic musicians, have begun to add a sound track not only to the often tangled workings of parallel programs but also to huge stores of data that comprise medical images, partial differential equations and even census figures. Last fall, for example, 35 followers of what has come to be called sonification, audification or, even less sonorously, data-driven sound descended on the Santa Fe Institute for the first International Conference on Auditory Display.

This tiny community of researchers is trying to bring another dimension to the communication between machine and user. Just as the ear can distinguish the oboe from the other instruments in a symphony orchestra, cybernetic audiophiles believe sound may complement or, in some cases, even surpass visual graphics.

Acoustic displays may reinforce visualization or add new information. A change in pitch, duration or loudness, for instance, corresponds to a variation in a value from the collected data. Some workers assert that existing technology severely overloads the visual. Yet, they point out, computers ignore the sense of hearing except for an occasional "wild eep" or bong from a neglected piece of software. Communicating information by means of an aural cue is a legacy of

Putting a Spin on Parasites

Adrian Parton used to spend his days pondering molecular biology. Then, in 1991, he heard a talk given by biophysicist Ronald Pethig of the University College of Wales, in which he described how electric fields can be used to manipulate particles. That set Parton, who works for Scientific Generics, a technology consulting firm in Cambridge, England, thinking about rotating fields.

A rotating electric field—easily produced from four electrodes—creates a torque on small particles that can make them spin if the frequency is right. Parton has shown that the effect can function as the basis for an extremely sensitive assay to detect microscopic parasites and even single biological molecules. Co-inventors Parton and Pethig call it the electrorotation assay.

Scientific Generics has now embarked on a joint venture with Severn Trent, a water utility in England, to develop and license equipment based on the principle. One of the first applications that will be pursued by the new company, Genera Technologies, is an apparatus to assess water-supply contamination by *Cryptosporidium,* a common waterborne parasite that caused an outbreak of sickness in Milwaukee, Wis., this past spring. Parton maintains that the assay could reliably detect other small beasties as well, such as *Giardia,* an intestinal parasite.

What is more, Parton says, the assay can outperform ELISA (which stands for enzyme-linked immunosorbent assay), a widely used laboratory technique for detecting small quantities of specific proteins. Genera plans to pursue several applications in the health care and food industries.

To detect *Cryptosporidium,* Parton first concentrates the particulates from a 1,000-liter sample of water into two microliters. He then adds microscopic beads coated with an antibody that binds to the organism's oocysts (the form found in contaminated water). The rotating field makes the oocysts bound to the beads spin. The number of spinning oocysts can be recorded electronically.

At low frequencies—around 1,000 hertz—live specimens rotate in the opposite direction to the applied field. Dead specimens, in contrast, rotate at higher frequencies but spin in the same direction as the field. The assay can thus assess the viability of *Cryptosporidium* as well as its numbers. "At frequencies of about a megahertz, you are looking at the ionic properties of the membrane," Pethig explains. "At still higher frequencies, you are looking inside the cell." A test that normally takes eight hours can be done in 30 minutes by electrorotation assay, Parton declares.

To detect targets smaller than oocysts, such as bacteria or protein molecules, Parton adds larger antibody-coated beads. The beads themselves are observed through a microscope. Any targets present bind to the beads, which makes them spin. Parton has discovered that in samples with very few targets, only a few beads rotate—but those that do so rotate very fast. That makes the assay exquisitely sensitive.

Parton says he has detected viral proteins in amounts as small as a billionth of a billionth of a gram. Commercial ELISAs now available are about a million times less sensitive than the levels that Parton claims. If beads are coated with nucleic acids instead of antibodies, Parton speculates, the technique could rival the polymerase chain reaction, currently the most sensitive test for particular sequences of genetic material. —*Tim Beardsley* the Geiger counter, sonar systems and the alarms that ring in the cockpits of jet fighter aircraft. Engineers in the field already see far beyond that simpleminded prompting. Sound may be the best means of depicting the ordering of several overlapping events, say Joan M. Francioni and Jay Alan Jackson of the University of Southwestern Louisiana.

The husband-and-wife team have produced scalelike tones that will let a parallel programmer tell when a processor becomes active or when it receives a message from another processor. Their choice of synthesized sounds even won plaudits from music aficionados at various conferences.

Other workers explore complementary strategies, in which sound and graphics reinforce each other. At the University of Massachusetts at Lowell, Stuart Smith, a professor of computer science, and his colleagues have combined visual and sound displays to exhibit variables ranging from infrared radiation from satellite imagery to heat flows in differential equations. Tiny zigzagging lines represent up to 20 variables graphically. Each connected segment, the zig or the zag, is a distinct variable, its value represented by the angle. Together they create an image whose texture varies across the screen like a field of rippling wheat.

The impression of an undulating topography is reinforced by tones that are reminiscent of Karl-Heinz Stockhausen's greatest hits. Like an auditor of one of the avant-garde German composer's works, the observer does not inspect each value separately but looks for clusters of similar sounds and visual cues. Staccato clicks and pops, translated from the brightness values for two merged magnetic resonance images, are supplemented by a crackling that sounds like an egg frying as a cursor is passed over an area showing a brain tumor.

Sound can pick out information that is difficult to discern in the image. Each person in a sampling of 1980 census data for scientists, engineers and technicians becomes a jagged line, whose segments represent marital status, level of education, sex and birthplace. These same variables can be represented by changes in sound. The category for sex, for example, is delineated by an electronic violin. A long, bowed note reveals a male; a high-pitched pizzicato note is the female. (The political correctness of the sundry tones has yet to be debated by technical standards groups.) As the mouse moves the cursor up the screen—the big wage earners reside at the top—the continuous tone from the bow gets well exercised. A short pluck in the upper corner, a lone woman, demonstrates how sound can pick out differences in the data that would be difficult to distinguish within the mass of visual cues.

Smith and other researchers in this auditory enclave are still asking themselves how much information can be clearly understood by listening to a data base. Training a listener to identify the single highly paid woman or heat-flow value may be a pointless pursuit. What is sought, instead, is the distinctive phrasing in the flow of synthesizer notes—a data melody—that will enable the listener to, say, locate a tumor.

The field is too young to have developed the conventions that would allow a set of characteristic notes to be recognized as cancerous tissues or heat flows. Smith's group recently received a $147,000 grant from the National Science Foundation to work with an experimental psychologist on determining how well the ear can hear the subtle changes in pitch, loudness and other sound traits within an aural display. "You can arrange a clever scheme for visualization and audification, but the question is how will people perceive it," Smith says.

Sound-display artists hope the ear will in some situations surpass the eye because of its superb ability to extract meaning from a noisy environment. Gregory Kramer of the Santa Fe Institute and W. Tecumseh Fitch of Brown University found that a group of university students who were assigned to role-play as medical doctors were able

Keeping the Sun Shining on British Technology

Michael Faraday, on being asked by a prime minister of the day what good were his researches into electricity, is said to have answered, "Sir, one day you will tax it." The successors to Faraday's interrogator may need to be reminded of the 19th-century physicist's vision. Margaret Thatcher, despite her professed admiration for Faraday and her chemistry degree, gave Adam Smith's invisible hand such a free rein during the 1980s that spending on research and development in Britain is now less, in relation to gross domestic product, than it is in the U.S., France, Switzerland, Germany or Japan.

A policy that prohibited government support for "near-market research and development" further battered Albion's technical prowess. The hope, of course, was that industry would pick up anything worthwhile that the government let drop. But the plan didn't work. The Confederation of British Industry (CBI) notes that U.K. patenting activity decreased by 6.4 percent between 1985 and 1990, for example, while it increased in Germany, the U.S. and Japan.

Britain's industrial decline now seems to have persuaded the prime minister, John Major, that it might be worth giving the Newtons, Watts and Darwins of the next century some encouragement. The government has for the past several months been conducting a full-scale policy review aimed at goading the scientific goose into laying more golden eggs. As in the U.S., where the Clinton administration has embraced technology policy, there is increasing talk of a government role in fostering innovation.

The man in charge of the exercise is William Waldegrave, a soft-spoken humanities graduate who enjoys the formal title of Chancellor of the Duchy of Lancaster. More meaningfully, he heads a new independent office of science and technology, the first cabinet minister with that portfolio since the 1960s. In June, Waldegrave was expected to publish his proposals for a complete shake-up of the British scientific enterprise.

His work is cut out for him: Waldegrave's office has received more than 800 pieces of advice. The most common complaint is one familiar in the U.S.: Britons, academics and officials like to say, are as good as they ever were at doing science, but industry fails to develop the commercial spin-offs. Industry charges that the government should be doing more to help. "We see no sign of strategic direction in the way resources are allocated," said Howard Davies, director general of the CBI, last November.

Interviewed in his Whitehall office, Waldegrave points out that counts of citations of published papers indicate that Britain is second only to the U.S. in scientific standing. "Britain has a powerful science base," he says. But, he admits, "we have continuing uncertainties about whether we're good at relating it to societal goals." Waldegrave, who has previous experience in industrial research as well as in science policy-making, contends that leaving decisions to industry has given British technical expertise "a certain random, spotty character."

Much in favor with industry is the suggestion that the U.K. adopt a R&D tax credit similar to that of the U.S. Yet Waldegrave is "pretty skeptical" about using to follow multiple aural variables on a computer—used to represent body temperature, heart rate, blood pressure and five other variables—better than they could on a graphic display.

Indeed, the next-generation computer operating system may enhance visual icons with sounds. When someone else accesses your computer, these "earcons," as they are whimsically called, might reproduce the tap-tap of approaching footsteps. "If you saw things through the same cone of vision as your narrow computer screen, you couldn't get a driver's license in the United States," says William Buxton, a professor of computer science at the University of Toronto and a researcher at Xerox's Palo Alto Research Center. "We're already overloading the visual channel. One key solution is off-loading appropriate things to the auditory channel."

Some things are just waiting to sound off. Chris Hayward, a seismologist at Southern Methodist University, wants to delve into what can be learned by speeding up and listening to the various subsonic shock waves of an earthquake. R. Kevin McCabe, a computer scientist at the National Aeronautics and Space Administration Ames Research Center in Mountain View, Calif., has written a program—a software "microphone"—that does digital sampling and playback of a computer simulation of acoustic data from a jet turbine blade. The days may be gone when a computer was meant to be seen and not heard. —*Gary Stix*

tax policy to stimulate economic activity. The government has also dismissed as too expensive proposals that it fund "Faraday Centres"—technology-transfer institutes modeled on the German Frauenhofer Institutes—to foster mixing among academic and industrial scientists.

Rather Waldegrave sees a case for improved direct support for "strategic" research in key areas, that is, work somewhere between completely unfocused "blue sky" investigation and product development. And he would like to see "a little more coordination and a little more structured interchange" between government, academia and industry.

OFFICE OF PUBLIC SERVICE AND SCIENCE

WILLIAM WALDEGRAVE is presiding over the first complete shake-up of British science in nearly two decades.

There are misgivings about the applied focus of the government's review in academic quarters. And Waldegrave insists he can promise no increases in support for civilian research. The $1.7 billion planned for the 1993–94 fiscal year was intended to be a modest boost, but scientific organizations argue that the devaluation of the pound turned it into a decrease.

The budget strictures mean there is likely to be no quick relief for impoverished scientists. Pay is almost unbelievably low by U.S. standards. A trainee researcher relying on the Science and Engineering Research Council for support, for example, has to survive the first three years after graduation on a stipend of less than $7,000 a year, a third of the national average wage. The government seems to have accepted the need for a remedy, and William D. P. Stewart, who is the government's chief scientific adviser, is sympathetic to proposals to make "blue sky" grants for young researchers easier to obtain. But, he warns, "any increases in support for research will have to be coupled with demonstrated increases in efficiency."

Britain will be looking for ways to improve efficiency in international "big science" projects as well, says Stewart, who is a botanist. He declares that the country cannot be party to the construction of expensive research facilities that stem from "national pride" (a dig at the U.S.'s Superconducting Super Collider). Rather he favors planned international collaborations.

Increasingly, British scientists say they are willing to recognize the financial realities of research, and some are even anxious to see their work commercialized. But dispelling the British economic malaise will be a tall order, even for a scientific magic wand. —*Tim Beardsley*

Practical Fractal

Mandelbrot's equations compress digital images

When IBM researcher Benoit B. Mandelbrot published *The Fractal Geometry of Nature* a decade ago, few imagined that the beautiful, infinitely detailed designs he called fractals might one day also improve television reception and help make pictures a part of everyday computing. Two who did were Michael F. Barnsley and Alan D. Sloan, mathematicians at the Georgia Institute of Technology. They reasoned that because images of the real world tend to consist of many complex patterns that recur at various sizes—in other words, fractals—there ought to be a way to translate pictures into fractal equations. Images so coded would require fewer data and thus less disk space to store and less time to transmit.

By 1987 the researchers had worked out enough details to patent their idea, garner $500,000 in start-up capital from Norwegian investors and form Iterated Systems in Norcross, Ga. But turning potential to profit proved difficult. Iterated found a formidable competitor in JPEG, an image compression technique developed by the International Standards Organization's Joint Photographic Experts Group. JPEG uses a well-understood mathematical procedure, the discrete cosine transform, to compress files by 90 to 95 percent—equivalent to compression ratios of 10:1 to 20:1. JPEG enjoys another advantage: because it is an industry standard, it is free.

The mathematics behind fractal compression, on the other hand, has only recently left the crib. "Relatively speaking, very little research has been done to date," Barnsley admits. "Two or three years ago, when people looked at my pictures, they said, 'This stuff looks terrible; it isn't going to work—you're dreaming.' But we bet on continual dramatic improvement, and we've seen it."

Indeed, fractal compression has at last begun to challenge JPEG in the three areas that matter: speed, quality and strength. To a certain extent, the three are mutually exclusive—stronger compression yields a smaller image file but takes longer and creates more artifacts. JPEG takes as long to decompress an image as to compress it, and it produces blocky "pixelation" as compression ratios rise above about 20:1.

Iterated takes a different tack. Its algorithm works by treating an image as a jigsaw puzzle composed of many overlapping pieces, some of which are similar. The software takes each piece and, using a fractal formula, transforms its shape, size and color until it matches another part of the picture. To do this for each piece of the puzzle takes some time; fractal compression is quite slow. But it produces a much smaller image file, which contains just the numbers needed to specify the mathematical relations between the pieces, and not those needed to actually draw each piece.

This approach has some unique advantages. Reconstructing the image from the numbers is relatively quick work: Iterated's video software can decompress tens of images each second without any special hardware. Quality depends on the accuracy of the numbers; giving the compression algorithm more time or a dedicated computer processor thus yields better pictures. And the artifacts generated at higher compression ratios add an impressionistic blur, rather than blockiness, to the image. What is more, because fractal images are encoded by equations, they have no inherent size—an attractive quality to publishers who want their products to work as well on a laptop screen as on a 21-inch monitor. Barnsley expects that in the long term, such resolution independence will prove irresistible.

So far, however, most of Iterated's customers have chosen fractal compression for its brute power. Microsoft licensed Iterated's software to compress the nearly 10,000 pictures it put on Encarta, its CD-ROM encyclopedia, because "fractal compression produced much more aesthetically pleasing results than JPEG at high compression ratios," says Jay Gibson, who manages the project. Racal Radio in Reading, England, put fractal compression into its PICTOR system for similar reasons. "We wanted to send data files over high-frequency radio, which is currently limited to some hundreds of bits per second," explains Peter W. Goddard, Racal's marketing manager. "Using fractals, we were able to get a 768-kilobyte picture file down to 10 kilobytes or lower—small enough to transmit in two to three minutes."

Other companies are attempting even higher compression ratios. Kalman Technologies used fractal compression to add video mail to its HydraWorks communications software. "We found compression ratios of about 600:1, which is what we need to send video mail cost-effectively across public telephone lines," says company president Steve Swift. Tayson Information president Peter A. Richardson says his company routinely obtained 100:1 compression when it added fractals to its education software, used by the University of Guelph in Ontario to create more than 500 undergraduate course modules. "We chose fractals because JPEG is pushing the limit of its capabilities," Richardson says. "Fractals can go and go and go."

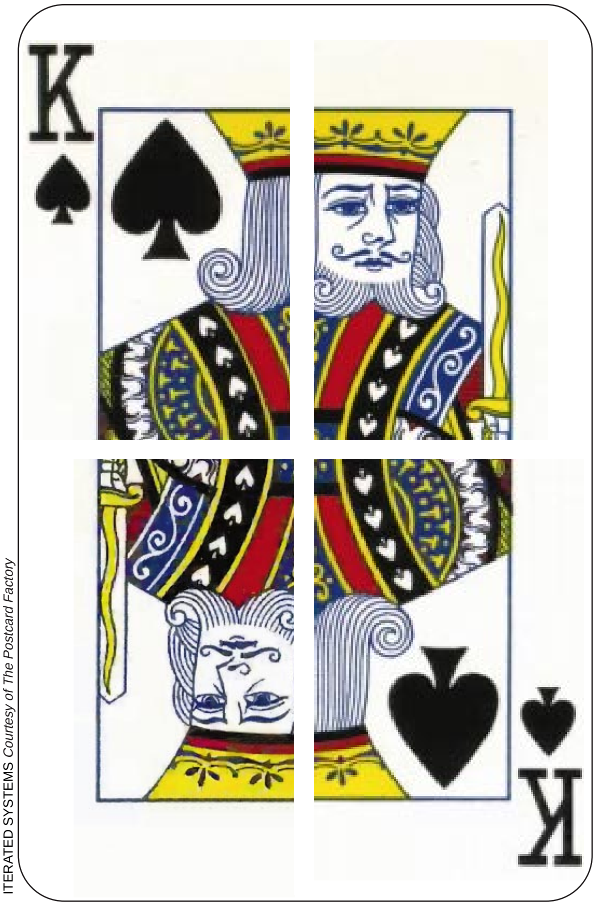
How far they can go is a matter of some dispute. Barnsley has claimed "effective" compression ratios of up to 2,500:1. But "sometimes people expand the image and then claim compression relative to the expanded image," notes Michael F. Shlesinger, who as director of physics at the Office of Naval Research funded Barnsley's early research. "It's just a trick." That trick has some mathematicians crying fraud and heaping scorn on Barnsley. Shlesinger attributes the controversy to a conflict of personalities. "Barnsley stopped playing the role of professor and started playing the role of entrepreneur," he explains. "He'll come to meetings and give a sales pitch for 45 minutes instead of a mathematics talk. This drives mathematicians up the wall."

Academic ire does not seem to be hindering Iterated's business, however. Five years and $18 million after its inception, Iterated finally started making money. "At the end of last year, we passed breakeven, and cash revenues have increased by over 10 percent a month each of the past six months," boasts Louisa Anson, a vice president. "Our goal is to go public in 1995." The company has grown to more than 56 employees (32 of whom work in research and development), offers half a dozen or so image compression products and is working to establish a foothold in the nascent computer video market.

Video will pose a greater challenge than still-image compression. The current video compression standard recently finalized by the Motion Picture Experts Group (MPEG) is well entrenched and near a second version. AT&T, Microsoft, Philips and numerous other giants have announced plans to sell MPEG-based video decoders within the year.

Barnsley cares little. If nothing else, he maintains, the resolution independence of fractals will triumph in the end. Iterated's long-term goal is to produce MPEG-compatible chips that use fractals to improve television reception and from there push into broadcasting. To that end, the company sought and received a $2-million grant from the National Institute of Standards and Technology to help it develop a fractal compression chip for high-definition television. "Once we get the reception circuitry in place, we're in a position to offer better transmission as well," Barnsley says. "It's kind of a Trojan-horse strategy."

In the meantime, Barnsley sees a niche in CD-ROM video for computers. MPEG requires special decompression chips to play video. Iterated has teamed up with TMM in Thousand Oaks, Calif., to develop software that plays broadcast-quality video straight from a CD-ROM. "We're going out full blast at Comdex [the computer trade show] with video," Barnsley beams. "By this fall we'll be a major player in the desktop computer video market." But first Barnsley and his mathematicians must finish their homework. *—W. Wayt Gibbs*



ITERATED SYSTEMS *Courtesy of The Postcard Factory*

FRACTAL COMPRESSION technique patented by Iterated Systems reduces data by translating a picture into fractal equations. The higher the compression, the more the image is distorted. **Clockwise from top left:** ***King of spades at original, 2.4-megabyte size, then compressed to 129, 63 and 31 kilobytes.***

THE ANALYTICAL ECONOMIST

Health Care without Perverse Incentives

To traditional economists, the U.S. health care system looks as if it had been designed to produce market failure. Customers have relatively little information about the product they are buying and must rely largely on sellers for advice. Sellers are paid according to how many procedures they perform, regardless of their effectiveness.

The price consumers pay for services bears only a distant connection to the amount they use. The middle and upper classes effectively buy their health care in prepaid annual chunks, either directly or as part of their compensation at work; the elderly and the poor have theirs paid for out of tax revenues. In all cases, the sum at stake when people decide whether to see a doctor is only a small fraction of the actual cost.

These perverse incentives have caused health care expenditures to soar and have led providers to exclude many people from the market. The U.S. now spends $800 billion a year to leave close to 40 million citizens without adequate care. This spending diverts resources from other issues on the national agenda and hurts the country's competitive position. For example, manufacturers complain that health insurance premiums add more than $2,000 to the price of an average automobile.

To combat these escalating costs, the Clinton administration has gathered more than 500 experts to create a new health care system. The central feature of the plan is supposed to be "managed competition," in which groups of consumers (either companies or individuals) bargain with medical consortia to purchase care at a reasonable price.

Many economists, however, are skeptical. They argue that the task force is ignoring fundamental economic principles, which put efficient, competitive health care markets in the same class as powdered unicorn horn. There is no way of arranging health care financing that does not lead to perverse incentives, contends Uwe E. Reinhardt of Princeton University: "We proved that years ago."

As long as doctors are paid on a fee-for-service basis, critics argue, they will overtreat—and fear of malpractice litigation tends to encourage unnecessary tests and procedures. As long as patients rely on insurance, they will demand more treatment. Economic theory suggests that forcing patients to cover a percentage of their medical bills would curb spending. But, as Henry J. Aaron of the Brookings Institution observes, if patients paid enough to make them really cost-conscious, a single bout of serious illness would spell financial ruin (as it does today for the uninsured).

Furthermore, Aaron points out, most health care dollars are spent during episodes of acute illness, when patients have little control over charges. Indeed, in another example of perversity, conventional limits on out-of-pocket costs may mean that patients have more financial incentive to postpone a checkup than to avoid a major operation.

Health maintenance organizations (HMOs), which offer to meet all of a consumer's medical needs for a fixed fee, have been cited as a major ingredient in the managed competition plan. But, as Reinhardt and others note, these groups create a different and opposing set of perverse incentives for physicians and patients. Doctors receive a fixed salary and so do best financially by offering as little care as possible; patients have paid up front and so profit (if that is the right term) by insisting on maximum treatment. The HMO itself has an incentive to sign up a maximum of subscribers and to minimize the time each one spends with a physician.

> *Efficient, competitive health care markets may be in the same class as powdered unicorn horn.*

Given this set of imperfect alternatives, critical economists say, the objective of health care reform must be to downplay the inherent flaws in this particular market rather than making futile attempts to eliminate them. But is this insight useful? Most of the suggestions they offer are modifications to the current status, not radical departures in the vein of, say, Great Britain or Canada.

Reinhardt proposes reducing perverse incentives by setting physicians' incomes according to the number of patients they treat—a little like setting up myriad microscopic HMOs—but allowing patients to choose freely among doctors, as HMOs do not. The reward for undertreatment will be balanced by the threat that patients will take their money elsewhere.

Although the ability to switch physicians is important, it is not enough: the cost of finding out whether one has the wrong doctor is quite high. So the next step is to monitor the quality of care and make that information available. Publicly disseminated physician profiles would help patients choose wisely, encouraging both competition and cost containment.

Indeed, Alan M. Garber of Stanford University believes medical quality assurance will eventually reshape the market: "We are faced with the problem of trimming fat from health care, and we are going to be cutting out treatments that do some good—but not very much and at high cost." Much of the recent debate over booming expenditures has focused on the question of whether high-technology medical interventions actually improve care.

But studies to determine when expensive medical procedures are vital and when they are placebos will carry their own costs. Administrative expenses already eat up nearly one fifth of every health care dollar in the U.S. In addition, as Victor R. Fuchs of Stanford comments, such numbers will do nothing to answer such crucial questions as "Are dying patients treated with compassion?" or "How long does a bedridden hospital patient lie in urine before someone responds to a call?" New measures of effectiveness may have to be developed if humane medicine is not to fall victim to cost-cutting.

Perhaps the most paradoxical problem in health care reform is that if managed competition or negotiated caps on medical spending do succeed in controlling costs, policymakers will be stifling one of the past decade's most potent sources of new jobs. Health care has been almost a textbook case of the transition from the smokestack industry to the service economy. "Every dollar spent on administration is a dollar of income to someone," Fuchs says. With almost one seventh of its gross national product dependent on health industries, the U.S. may be in a classic double bind.

—Paul Wallich and Marguerite Holloway

The Topological Dressmaker

The floor was littered with bits of paper pattern, pins and snippets of cloth. Jemima Duddlepuck whistled as she worked on her new summer dress, while in a corner her dog, Pucci—who was rather lazy and very fat—snored happily. It was the first time Jemima had ever tried to make a dress with a lining. All that remained was to sew the two together, so that the seams ended up inside, between lining and dress, invisible from the outside.

She looked at the instructions. "Turn the lining inside out. Slip it over the dress. Machine around the neck and armholes. Turn the completed dress the right side out." Ever so gingerly, she machined around the armholes and the neck. The hems, of course, were left unattached. Done! Oh, it was going to be *really* nice. Jemima began to turn it inside out.

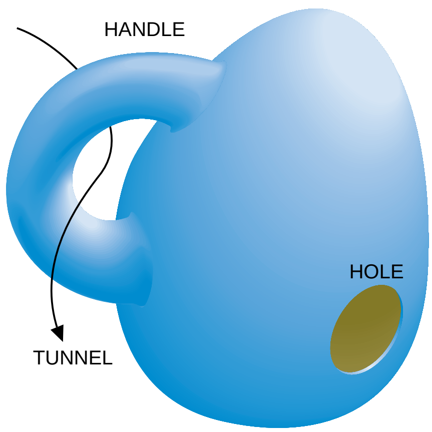
Funny. Was something stuck? No, if she turned it back to the start, then everything seemed okay. But that final twist eluded her. Half an hour later she hurled the balled-up wad of material into the corner of the room and burst into tears. All that work for nothing. Then, having vented her frustration, she sat down to think the matter through. Jemima believed in learning from mistakes.

Had she made some parts the wrong shape? No, that wasn't likely. If you couldn't turn a slightly distorted dress inside out, then you couldn't turn a perfectly made one inside out either. Whether or not you could turn something inside out was a topological question—the answer didn't change if the object was continuously deformed. That meant two things. First, there was something wrong with the instructions in the pattern book—because whatever mistakes she'd made, they had to be very small ones. Second, there had to be a *topological* reason why the instructions were wrong.

What could she remember about topology? Well, there was the idea of topological equivalence. Two objects are "the same," or topologically equivalent, if one can be deformed continuously into the other. You can bend them and stretch them and shrink them and twist them, but you cannot cut them or tear them or glue bits together that weren't glued together to begin with. That was the basic idea—but what kind of things did topologists study? There was a lot of stuff about Klein bottles and Möbius strips, surfaces with only one side. But that wasn't likely to be relevant to problems about things with *two* sides. Then there was some other stuff about surfaces with holes. That was more likely to be useful.

She recalled that "hole" is a rather ambiguous word in topology. For instance, people usually say that a donut—or torus—has a hole in it. But the hole isn't really *in* the donut at all. It's just a place where the donut isn't. Whereas if you take a car inner tube—also a torus—and cut out a patch of rubber, then you really do get a hole.

All right, she thought, I'll call the kind of hole you cut out with scissors a hole,



Some topological terminology

and the kind where the surface just isn't I'll call a tunnel. One way to create a tunnel is to add a handle to a simpler surface [*see illustration above*]. By adding handles one at a time to a sphere, you first get a torus, with one tunnel through the middle. And then there's the two-handled torus, the three-handled torus and so on. She knew these were often called the two-holed torus and so on, but she wasn't keen to use the "hole" terminology without being very careful indeed what she meant.

Now, the instructions were to turn the lining inside out, so that its outside was on the inside; machine the neck and armholes; and then turn both lining and dress inside out so that the inside of the lining was next to the inside of the dress—no, the outside of the inside of the…no, the inside-out lin—. Oh, heck. Terminology for sides was going to be at least as much a problem as terminology for holes.

Start again. The dress has two sides: the outside, which is what everyone sees when you wear it, and the inside, which is what they don't see. The lining also has an inside and outside, but those are confusing words. So call the side of the lining that fits next to the skin the skinside and the other side of it the skirtside because it goes next to the dress. Well, as mnemonics go, those weren't too bad. It reminded her of George A. Strong's parody of *Hiawatha:*

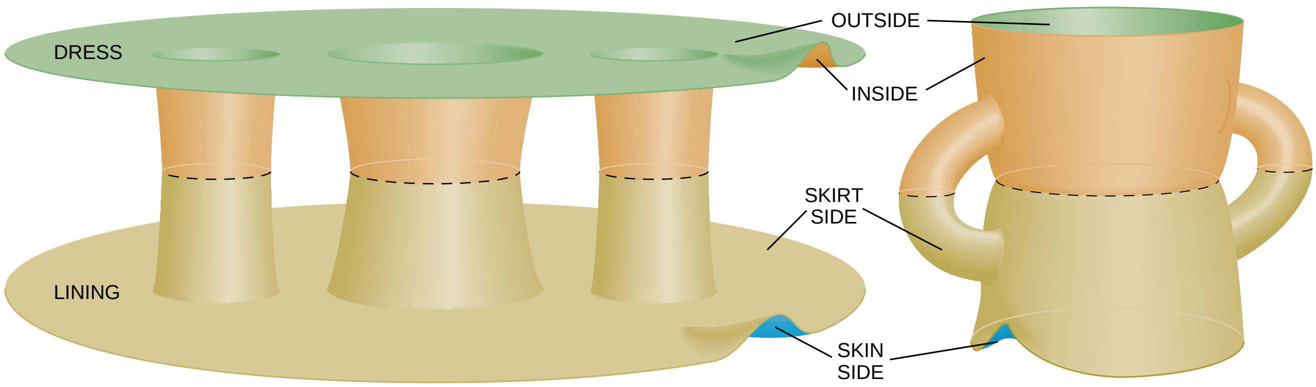
> When he killed the Mudjokivis,
> Of the skin he made him mittens,
> Made them with the fur side inside,
> Made them with the skin side outside,
> He, to get the warm side inside,
> Put the inside skin side outside.

So the sides of the material, reading outward, have to end up in this order: skinside, skirtside, inside, outside. And they had to be machined so that the stitches around the armholes and the neck were hidden between skirtside and inside.

Topologically, both the dress and lining were equivalent to disks with three holes. The rim of each disk was the hem, and the holes were the two armholes and the neckhole. What the pattern book said was to place the disks so that the skinside was next to the outside; machine all three holes so that the stitches were on the skirtside/inside surfaces [*see top left illustration on opposite page*] and then turn the whole ensemble inside out.

So the question was, what happened? Well, she thought, you can deform the whole ensemble topologically until it becomes a two-handled torus with two holes cut in it. (To see why, imagine sliding both ends of the two armhole tubes onto the neckhole tube and then bending the dress hem up and the lining hem down and shrinking them [*see top right illustration on opposite page*].) So the problem becomes: Can you turn a two-handled torus inside out if there are two holes in it?

It was kind of hard to think about, so Jemima started with something simpler.

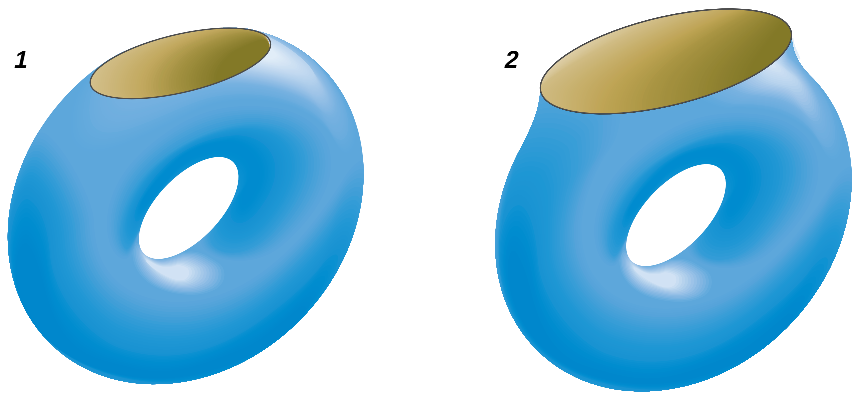


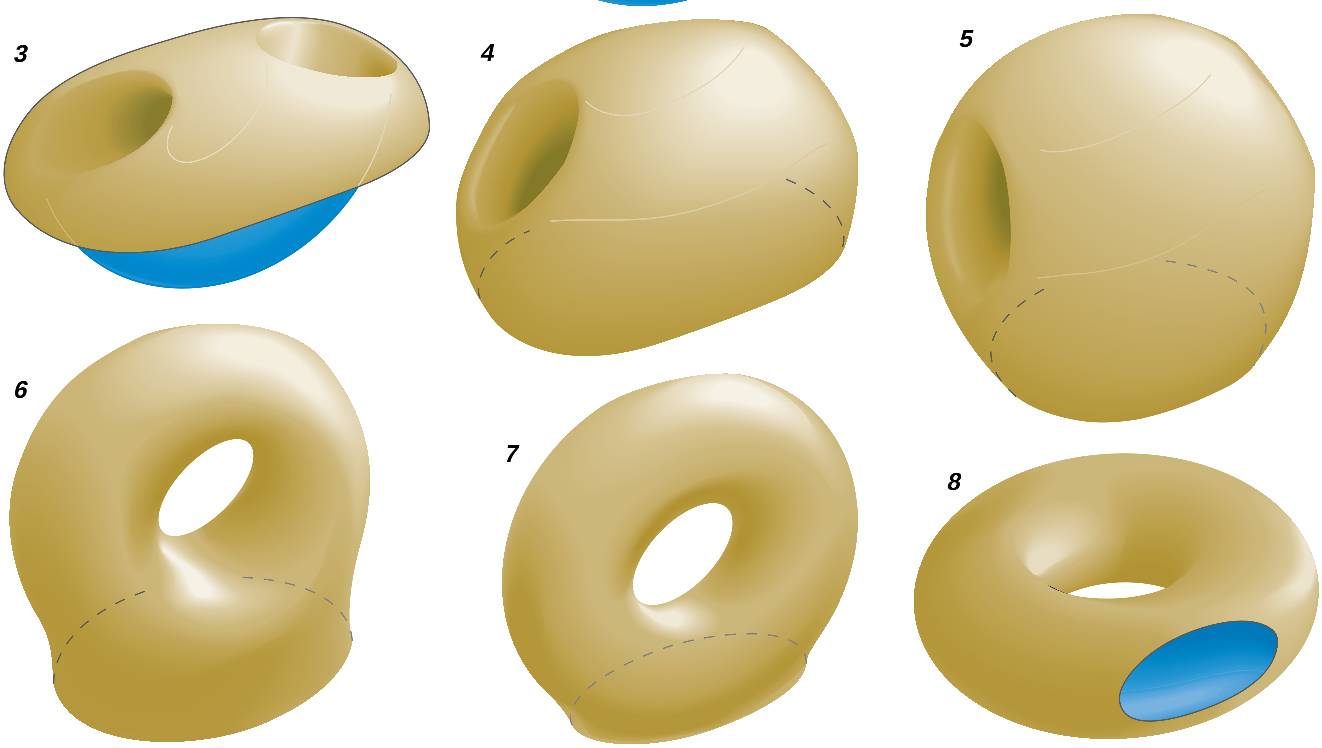
DRESS PLUS LINING is topologically equivalent to two disks, each with three holes, sewn together along three edges **(left).** ***This configuration in turn is topologically equivalent to a two-handled torus with two holes, hems of dress and lining*** **(right).**

Can you turn an ordinary torus inside out when *one* hole is cut in it? And the answer, of course, was yes. The idea was to flip the entire thing through the hole, but then the handle was like a tunnel bored into a sphere rather than a handle stuck on the outside. But you could imagine sticking a finger into the tunnel and straightening it. Then the part of the surface "outside" the tunnel is a rather distorted handle, which can be tidied up to look like a neater one. Then you got something that looked just like a torus again [*see illustration below*]. So *topologically,* there was no difficulty.

She realized she could play the same game on the two-handled torus, too. She sketched a series of pictures of the process. First, turn the whole thing inside out through its holes, rather like reversing a sock (or in this case an ankle warmer because of the two holes). Then the two handles on the outside became two tunnels on the inside [*see illustration on next page*]. But then you can stick your finger into each in turn and pull it out to create a handle on the outside. And then all you have to do is twist each handle around, and you end up with the original ensemble, but turned inside out.

Hmm. You *could* do it topologically, then. Strange, because when she tried it with an actual dress, it didn't work.





How to turn a torus inside out through one hole

Why not? Dresses, unlike topological spaces, can't be stretched or shrunk. They can be crumpled up and twisted around, though. It was possible that the material nature of the dress changed the answer, but she had a feeling that something both simpler and more fundamental was involved. She picked up the ruined mess from where she had flung it and smoothed it out. She could wriggle her arm between the various surfaces, and it really did seem as if there were three handles. But near the armholes and neckhole everything was confused, as if the lining and the dress were getting mixed up. There were three handles, all right, but they just weren't in the right places.

Aha! She'd forgotten to think about the seams. They had to end up in the right places, neatly separating dress from lining. But did they?

She drew the seams on her pictures. No, they didn't. When the surface was stretched, to create the handles from the tunnels, and then twisted to put the handles in the right place, all three seams ended up in completely the wrong places. (E. C. Zeeman of the University of Oxford has proved that there is no way to make them end up in the right places.) So *that* was the topological obstruction to turning the ensemble inside out. You couldn't do it if the seams had to end up where they started.

And that gave her another idea. Maybe the ensemble wasn't ruined after all. She could unpick the stitches and try to find some other method of making the seams. She would be on her own, though—obviously, the instructions on the pattern were crazy.

Jemima unpicked one armhole, then the other. She was about to start on the neck when she had another thought. Topologically, each hole was just as good as the others. You had to start by machining *some* seam. The neck was as good a place to start as any.

Let's experiment, she mused. She pulled the lining inside the dress, skinside inside, then skirtside next to inside of dress, then the outside of the dress—well, outside.

Great! She even found that she could poke the arms of the lining down the arms of the dress, just where they ought to end up. The only trouble was, she couldn't machine around the seam; the stitches would be on the outside.

So turn just the arm inside out. She pulled the arm of the lining inside the dress, turning it inside out like a sock. Then she did the same with the arm of the dress. She burrowed into the ensemble from the hem end, flashlight in hand. It certainly looked as if the arm could be restored to its correct position. All she had to do was push the sewing machine up inside the ensemble and then—

Stupid. Sheepishly, she backed out of the heap of crumpled cloth. She reached in through the neck and pulled the sleeve out into the open. She ran the seam through the machine, then turned it inside out and back to its proper position. Now the dress really *was* finished. She tried it on and pirouetted in front of the mirror. Brilliant!

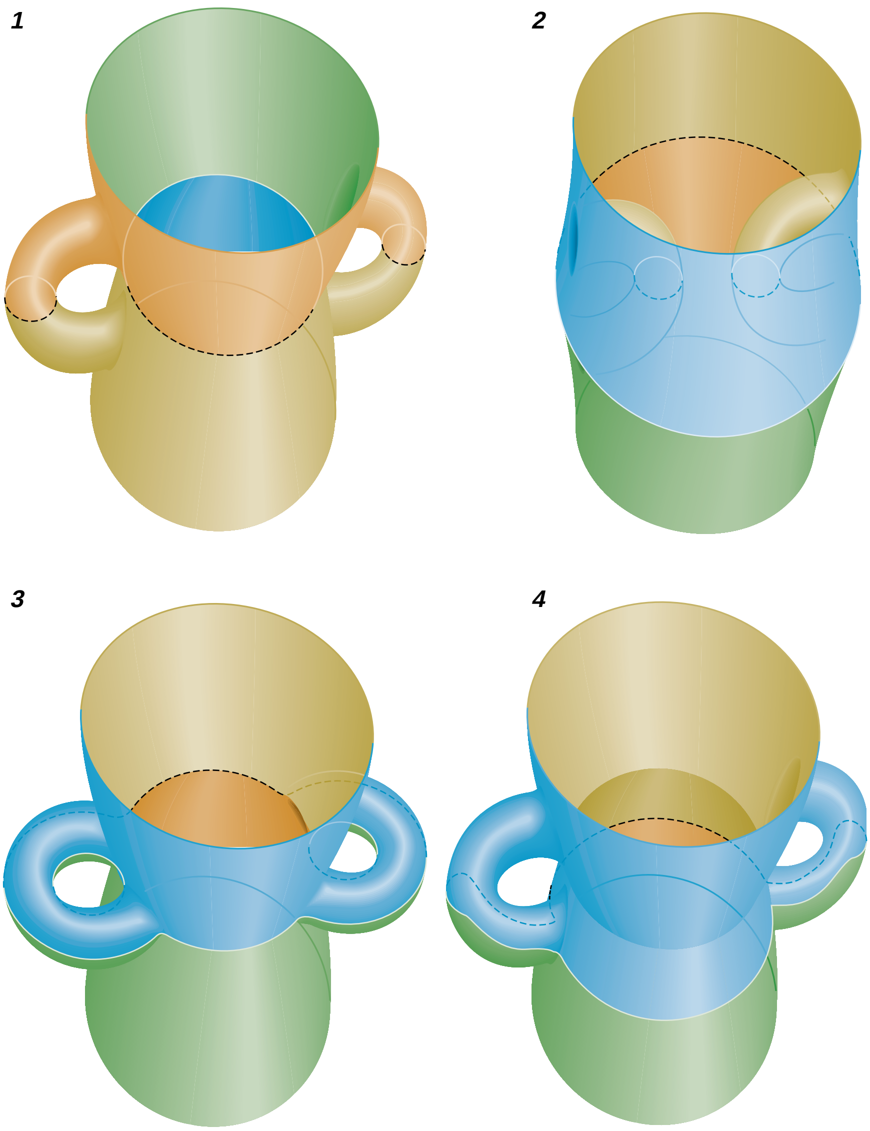
I ought to write a book about this, she reflected. *Topology for Tailors.* I could throw in all the party tricks, like taking off your vest without removing your jacket.

By now her mind was in overdrive. Suppose you were a Martian with 17 arms. How many hems would be needed to make a Martian ensemble, dress plus lining? Seventeen? No, just two. Use the same trick of machining the neck only and then dealing with each arm in turn by pulling it through the neckhole, inside out. Unfortunately, there weren't any Martians. So she wouldn't get to try it out—"Pucci, stop that, your nose is all wet!" She pushed the little dog gently away.

Say, Pucci had four "arms," and its tail could poke out through the hems. Jemima took the dog's measurements, tut-tutting at the figures for the waistline. She cut shapes from remnants of the cloth, whistled, sewed, twisted and turned. Yes, even with four "arm" holes, the method worked fine. Now to see if the dress fitted the dog.

It *fitted,* in the sense that the measurements were okay. But fat dog into small hole won't go.

In the real world, there are nontopological constraints.



TO TURN the ensemble inside out, reverse the whole thing. Pull the tunnels out to form handles. Twist the handles into correct position. Note, however, where seams end up.

BOOK REVIEW by Gerard Piel

Can History Stop Repeating Itself?

PREPARING FOR THE TWENTY-FIRST CENTURY, by Paul Kennedy. Random House, 1993 ($25).

In 1988 Paul Kennedy welcomed the Bush administration into office with the advice that the U.S. had begun to follow the British Empire into the inevitable second phase of *The Rise and Fall of the Great Powers,* as his popular book was entitled. Now the J. Richardson Dilworth Professor of History at Yale University welcomes the Clinton administration with the word that there is more to history than history.

In *Preparing for the Twenty-First Century,* he asserts that political leaders must reckon with three crises approaching from outside of history as it is usually told. Kennedy sees them, inextricably interconnected, as the current doubling of the world population, the peril in which the crowding human presence places the global ecosystem and the disruption of social institutions as well as the environment by technologies that are taking over life functions, including those of the human nervous system.

CLIFFORD HAUSNER *Tony Stone Images*

EXTENSIVE SLUMS in Rio de Janeiro kept the 1992 United Nations Conference on Environment and Development conscious of the world's most serious problems. (The photograph does not appear in the book under review.)

It augurs well that a historian should thus bring into his reckoning the long-term continuities of the human experience. With better comprehension of those realities, historians and the political leadership they seek to instruct might begin to find ways to bring history to stop repeating itself.

In part one of his book, Kennedy gives the current exigencies a compelling presentation. He states the predicament of population growth in a single, trenchant paragraph. Humankind numbered one billion at the end of the 18th century, when Thomas Malthus first estimated its size and made his familiar prediction of the cruel consequences to follow from what he perceived to be its inexorable and accelerating growth. In fulfillment of Malthusian misgivings, the number of the earth's inhabitants doubled in the next century and then doubled again in the next half century. By 1990 it had increased to 5.3 billion.

Having given Malthus his due, Kennedy points to a new phenomenon in human population dynamics: increases have all but halted in the industrial countries. "Even among today's fast-expanding populations of the developing world," Kennedy says, "demographers expect average family sizes to decline in the future, as urbanization and other factors cause a demographic transition and numbers begin to stabilize."

The demographic transition has a longer existence in human experience than this passing mention suggests. Demographers see global population moving from the near-zero growth that persisted up to 1600 at high death and birth rates, with life expectancy as short as 25 years, back again to near-zero growth at low death and birth rates, with life expectancy exceeding 75 years. The population explosion began in the 17th century in the now industrial countries. Representing one quarter of the world's inhabitants today, these nations have nearly completed the demographic transition. The leading edge of industrial revolution—sanitation and vaccination, mass literacy, the green revolution in agriculture—has crossed into the developing countries. In this trend, demographers find grounds to forecast stabilization of the population at a number technology and the earth's finite resources can sustain in civilized comfort.

"But that is decades away—*even if* those forecasts are correct," Kennedy concludes. "Before we reach what is termed 'global replacement fertility' levels, which United Nations authorities believe may occur around 2045, the population supertanker will have moved a long way."

It must indeed make a difference, as Kennedy argues, whether human beings stabilize their numbers at 10 billion or 20 billion. This, not the nightmare increase to terminal misery, is the population crisis.

The human presence already perturbs the life cycles of the planetary ecosystem. The industrial revolution, by making all the citizens in some nations rich and by lengthening the life expectancy of people even in the poorest nations, has amplified that influence. In their different ways, the rich, by their appetites, and the poor, in their increasing number, devastate the local, regional and global environment.

Kennedy gives first priority to the threat of global warming. (The evident effects of the thinning of the ozone layer on photosynthesis may soon give that threat higher priority.) Observation has established a measurable increase in the current concentration of carbon dioxide in the atmosphere; that is the result principally of the fourfold increase since 1950 in the combustion of fossil fuels. The geologic record shows the average ambient global temperature rising and falling in correlation with change in that concentration. Whatever scientific uncertainty qualifies the present threat, the rise in sea level and rearrangement of the climate map consequent to even a small increase in global temperature in the past should be compelling action now.

What is more, the economic development of the poor countries, crucial to the timing of their demographic transition and so to the size of the ultimate population, portends another fourfold multiplication of the output of energy. Without alternatives to fossil fuels, the environment appears in Kennedy's prognosis to be damned if they do and damned if they do not make the demographic transition.

Kennedy finds technology posing equal dilemma to the social arrangements of humankind. "Dematerialized" manufacturing technologies, no longer bound to the source of their raw materials, allow transnational corporations to seize the comparative advantage of poverty and build their factories in low-wage preindustrial countries—or staff them with robots at home. The extended, collective human nervous system arrayed in the high-velocity, high-capacity global communications network conducts 24-hour-a-day trading and speculation in national currencies, "90 percent of [it]... unrelated to trade or capital investment." In the borderless world economy the waning of the national state "erodes a people's capacity to control its own affairs."

Along with factory ghost towns, bankrupted by the "transnationalization of manufacturing," the U.S. has prairie ghost towns depopulated in the "trivialization" of agriculture by new technology from the life sciences. The same biotechnology that ensures increase of food supply in pace with the growing numbers of people may deprive preindustrial villagers of their livelihoods by "significant relocation of agricultural production (or its substitutes) out of the developing world" or by "in vitro production" of food in situ.

On territory more familiar to historians in part two, Kennedy conducts a comprehensive survey of nations and regions, extrapolating from their present situations and trends possible courses for each into the 21st century. He assesses their prospective ups and downs and their chances to arrive at mid-century as winners or losers. His assessments hold no surprises.

Japan is thus a likely winner. That is qualified in large part by the degree to which dematerialized technology may incorporate the Japanese world-bestriding industrial enterprises into the planetary family of transnationals. At present, Japanese corporations remain devoutly Japanese. For the celebrated "tigers of East Asia," Kennedy makes an almost equally hopeful but separate appraisal. Here he fails to reckon with the reality of the Greater East Asia Co-Prosperity Sphere. The tigers and still bigger new partners, Indonesia, Malaysia and Thailand, run huge deficits in their trade with Japan; they offset them by a combined surplus in their trade with the U.S. that is equal to Japan's. Surely, the future of this empire, now in such good working order, strongly conditions the future of each of its members, starting with Japan.

India and China, with one third of the world's people, command consideration together. Population growth reduces their impressive achievements in economic development to small per capita gains, India's smaller than China's. Equity in distribution of gains, more so in China than in India, has advanced both nations into the demographic transition, again China ahead of India. Together they are bound to make huge additions to the ultimate world population and still huger contributions to the concentration of carbon dioxide in the atmosphere. Economic change will put their unity as nations under stress. For individual Indians and Chinese, life will be hard but less brutish and short.

In assessment close to prophesy, Kennedy sees the calamity of Africa below the Sahara extending without remission into the 21st century. Second only to that of Africa, the poverty of Latin America promises to resist whatever development its countries achieve. In this connection, Kennedy does not give enough weight to the inequality, amounting to inequity, in income distribution prevailing in the region: the top 20 percent income group in Brazil keeps 62.6 percent of the gross national product and allots 2.4 percent of it to the bottom 20 percent.

Hardest to decipher, Kennedy decides, are the futures of the erstwhile U.S.S.R. and its former buffer states in Eastern Europe. Russia is bearing the brunt of the Soviet "triple crisis"—the economic incompetence of a lopsided industrial system, the uncertain legitimacy of the successor political powers and the rush of ethnic minorities to self-determination. However it weathers the near term, Russia faces the long-term ethnic "revenge" of population explosion in the newly sovereign Siberian republics. (They, as well as Kennedy, may call it "revenge," but they owe their population explosions to Russian promotion of economic development in the Soviet internal Third World.) Poland, the former Czechoslovakia and Hungary are prospective "fast adjusters," with economies recovering to 1989 levels as early as next year. Even the "slow adjusters"—Romania and the Balkans—get there ahead of Russia and the other fragments of the U.S.S.R.

The alternative futures of the Western European countries hang on their progress toward economic and political community. Concentration of economic power in their transnationals clouds the political scene: it hardens national borders even as it transcends them. Europe must continue under pressure of migration from poverty in West Asia and North Africa, unrelieved by fundamentalist Muslim rebellion. Despite all uncertainties, science, technology and civilization seem to hold out the good life to the next generations of Europeans.

Five years have given Kennedy no reason to adjust the trajectory on which he pointed the U.S. in his earlier book. He stoutly maintains his "declinism" against the protests that have come meanwhile from die-hard supply-side publicists. The U.S. has a combined public, business and household debt approaching 200 percent of the world's biggest GNP. It borrows $100 billion a year from the world economy. It loses one fifth of its children from school in their high school years. The number of candidates for advanced degrees in its universities is declining. Homicide rates per capita are five times the worst found in other industrial countries. The U.S. crowds more than a million convicts into its prisons, among them 40 times as many black prisoners, proportionately, as South Africa at the peak of apartheid. It holds family incomes at mid-1970s levels, even as the average compensation of the managers of failing industrial enterprises ascends from 40 to 90 times their payroll averages. It celebrates gains in "productivity" calculated from increases in applications for unemployment insurance. The U.S. can have no plan for the 21st century "without becoming a different kind of country." For that, short of a "sufficient shock to complacency," the prospect is small. "That, in turn, would require leadership very different from the sort demonstrated by recent incumbents of the White House, whether it concerned domestic deficits or global population and environmental issues."

On those issues that intrude from the world outside, Kennedy has little help to offer the leaders of the U.S. and the other industrial nations. He is misled on his venture into unfamiliar territory. The 21st century is not irrevocably committed to fossil fuels and climatic change. Investment now in promising alternative energy conversion cycles can meet the fourfold increase in energy demand without further opaquing of the carbon dioxide window. The capture of sunlight by photovoltaic cells and storage of the energy in hydrogen are al-

SCIENTIFIC AMERICAN

CORRESPONDENCE

Reprints are available; to order, write Reprint Department, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 10017-1111 or fax inquiries to (212) 355-0408.

Back issues: $6 each ($7 outside U.S.) prepaid. Most numbers available. Credit card (Mastercard/Visa) orders for two or more issues accepted. To order, fax (212) 355-0408.

Index of articles since 1948 available in electronic format. Write SciDex™, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 10017-1111, or fax (212) 355-0408.

Photocopying rights are hereby granted by Scientific American, Inc., to libraries and others registered with the Copyright Clearance Center (CCC) to photocopy articles in this issue of SCIENTIFIC AMERICAN for the fee of $3.00 per copy of each article plus $0.50 per page. Such clearance does not extend to the photocopying of articles for promotion or other commercial purposes. Correspondence and payment should be addressed to Copyright Clearance Center, Inc., 27 Congress Street, Salem, Mass. 01970. Specify CCC Reference Number ISSN 0036-8733/93. $3.00 + 0.50.

Editorial correspondence should be addressed to The Editors, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 10017-1111. Manuscripts are submitted at the authors' risk and will not be returned unless accompanied by postage.

Advertising correspondence should be addressed to Advertising Manager, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 10017-1111, or fax (212) 754-1138.

Subscription correspondence should be addressed to Subscription Manager, SCIENTIFIC AMERICAN, P.O. Box 3187, Harlan, IA. 51537. The date of the last issue of your subscription appears on each month's mailing label. For change of address notify us at least four weeks in advance. Please send your old address (mailing label, if possible) and your new address.

ready in pilot-plant development with government funding in Japan and Germany. Biotech "refineries" of synthetic nourishment—feedstocks not specified—are unlikely to obviate farmlands and villagers to cultivate them.

With respect to the demographic transition in the preindustrial countries and the economic assistance needed to encourage it, Kennedy has even less help to offer. He settles for distribution of contraceptives to the natives and, in an improvement on that condescension, the encouragement of female literacy. It is true that people in the developing countries are motivated to control fertility at much lower incomes than those at which contraception became widespread in the industrial countries. Thus, they are making the demographic transition in a much shorter time. An extensive literature on economic development supports the necessity, however, of appropriate physical and cultural setting for promotion of female literacy and motivation of fertility control: potable if not running water, a modicum of sanitation and schoolhouses, to start with.

Kennedy refers to a "UN Conference on Trade and Development" and "$125 billion a year to pay for new environmental programs…swiftly trimmed to a request for an additional $5 to $10 billion a year, in recognition of today's political and economic realities." Whatever happened at that UNCTAD meeting in Paris in 1990, that $125 billion stays there on the green-baize table at which the rich and poor countries are negotiating the future of the planet and its inhabitants. At the U.N. Conference on Environment and Development in Rio de Janeiro in June 1992, the then 172 rich and poor member states adopted Agenda 21. In preparation for the 21st century, this calls for annual investment in the economic development of the poor countries at the rate of $600 billion. The poor nations are to supply nearly $500 billion in the form of labor and natural resources. The $125 billion is the share of the rich countries. It is specified in the form of the catalytic "technology transfer" necessary to put those people and resources to work.

Agenda 21 stands as the first international agreement premised on the demographic transition. Its great objective is to hasten the arrival of the human population at a smaller ultimate size. While it is a nonbinding agreement, it does set out a "global consensus and political commitment at the highest levels" on what needs to be done about population, environment and development, comprehended in a program of sustainable development. By action of the U.N. General Assembly in December, Agenda 21 is now the agenda of the U.N., at the very least an agenda for ongoing debate in the General Assembly.

Implementation of Agenda 21 relies, of course, on the forthcoming of that $125 billion a year. As Kennedy notes, the Nordic countries (and Canada) are laying out 0.7 percent of their GNP in economic assistance to the developing countries. They are only making good on a promise in which all the industrial countries joined 20 years ago. In 1961, on a motion made by John F. Kennedy, they had promised 1.0 percent of their GNP to make the 1960s the "Decade of Development." When the 1960s went into history as the decade of disappointment for the preindustrial countries, the industrial countries committed to the more "realistic" figure of 0.7 percent.

By no coincidence, that is precisely the percentage of the GNP of the industrial world that yields $125 billion. At Rio, the rich countries, with the exception of the U.S., promised to bring their outlays for economic assistance up to the 0.7 percent "as soon as possible." The European Community and Japan were ready to make a down payment on their promise. They refrained out of diplomatic courtesy to the Bush reelection campaign and are waiting for the Clinton administration to settle into office.

If Agenda 21 has been so little celebrated in the U.S. mass media as to make this mention of it the first that has called it to the reader's attention, more will be heard of it during this last decade of the 20th century. Agenda 21 now has its own undersecretary general for sustainable development installed at U.N. headquarters in New York City and a determined caucus of developing countries in the General Assembly. It is gathering the support, moreover, of the citizen movements around the world that put environmental protection agencies into nearly every national government during the past two decades. At Rio, in their own Global Forum, 20,000 environmental zealots learned that poverty connects the integrity of the environment to development.

Anxious to create jobs, especially in the smokestack industries, President Bill Clinton may soon find it desirable to seek to increase the U.S. annual outlay to "foreign aid" above the present paltry and misdirected 0.2 percent of the country's GNP. His vice president is on record advocating a "global Marshall plan."

More sympathetic acquaintance with the relevant "political and economic realities," as well as with continuities of human experience less familiar to him, might have encouraged Paul Kennedy to urge bolder action.

GERARD PIEL is chairman emeritus of Scientific American.

Science: How Much Is Enough?

The military defense of the U.S. has been the best articulated motivation for the support of basic science. Also important has been a widespread belief that scientific leadership would translate into both economic power and a higher standard of living. The disappearance of the threat from the Soviet Union has undermined one of these rationales. The slow growth of the American economy, contrasted with the rapid progress of Japan, which is not a world leader in academic science, has eroded the other.

Does basic research have something to offer when economic progress and improving the environment have displaced the military threat on the nation's agenda? If so, how much basic research is needed?

The history of the computer answers the first question. The transistor could not have been invented without years of basic research that built a quantum mechanical understanding of crystalline solids. Without that work neither the device itself nor the subsequent gains in computing power accompanied by declining cost per circuit element would have been possible. We would not have the computer industry as we know it.

The research in solid-state physics that preceded the invention of the transistor exemplifies one kind of relation between science and economic progress. In this case, it is the emergence of useful technology from research pursued without a visible industrial goal. At the next stage are areas of science such as high-temperature superconductivity that have not generated new industries but do have discernible economic potential. Then there are new science-based industries closely dependent on fundamental research that already yield products. Biotechnology is one. It is not a large industry today, but it has enormous potential. This kind of infant industry, moving at a fast technical and competitive pace, requires continuous input of new science to maintain its expansion.

At the more developed stage reached by the semiconductor and pharmaceutical industries, technical directions have stabilized. In such industries, corporate R&D laboratories meet much of the scientific needs. Then there are industries in which technological change itself is slow. For these the tie to current science is correspondingly weak. Automobile manufacturing is an example of an industry that requires only the science that is available to everyone.

For industries of either of these last two types, ones that are like semiconductors or like automobiles, productivity gains and market dominance come, for the most part, from improving design and production processes, perceptive marketing strategies and managing people to benefit from their most creative capabilities. In some of these industries, we observe that Japan, in spite of its lack of prominence in academic science, has become a global power.

For fledgling science-based industries, then, the linkage to research must be close and continuous. Industries that are already well on their way can be sustained by the work of their own laboratories. For mature industries, the main contribution of science may be the graduates who can use the knowledge available to all.

Clearly, science does have an economic role. How do we decide what science to support and at what level to support it?

Our proposition is a simple one: all major fields of U.S science should be funded at a level that will keep them among the world's leaders. Being one of the leaders will enable us to be among the first to participate in the birth of new industries. This basic economic rationale must apply to all disciplines, because we do not know where the next significant event will take place. Moving quickly when something exciting does happen in a field may even call for increasing the support of that scientific area. Action may be necessary downstream as well.

We are suggesting that the economic criterion for deciding how much to support American science is a comparative one. Can comparisons by disciplines actually be carried out? The many studies conducted by the National Research Council, the opinions of researchers in the field and the number of first-rate scientists in a particular area all provide useful indicators.

We believe the goal of having a world-class scientific enterprise can be obtained without increasing the federal budget. The reason is that in the U.S. most fields of science probably already satisfy our test.

For those sciences that directly support the infant industries we have described, we propose that the U.S. should maintain a definite advantage. We might expect, for example, to see strong support for those parts of molecular and cellular biology and immunology that assure a primary position in therapeutic and diagnostic technology and in agricultural biotechnology. We would expect to see such support for condensed-matter physics, for chemistry and for the material sciences that contribute to the development of products based on high-temperature superconductivity.

None of these considerations should exclude funding based on other objectives and values. They might include contributions to health, to the environment or to other societal goals or a decision to emphasize a field because of its intellectual excitement. The economic rationale based on the comparative world-class criterion, however, should put a firm floor under the support of basic science.

In some cases, for example, space science, we may be spending billions more than anyone else. Funds from such fields should be redirected to areas in which our advantage is less overwhelming. In areas of inquiry—particle physics, for one—that depend on a small number of expensive, experimental installations, our standing could probably be maintained more easily by dealing with the complexities of international cooperation than by absorbing the enormous expense of an unneeded rivalry for world leadership.

We believe this approach would lead to a more predictable, stable scientific environment in which the funding could be determined on a long-term basis. It would help science to be a more attractive career for both the established and the young scientist. It would assure that our leading scientists would be well funded and that our country would be in a position to benefit economically from their work.

RALPH GOMORY is president and HIRSH COHEN is program officer for industrial studies at the Alfred P. Sloan Foundation in New York City.