# SCIENTIFIC AMERICAN

OCTOBER 1993
$3.95

*Is Los Angeles winning the war on smog?*

*Computers that mimic damaged brains.*

*Vast lava flows that reshaped the earth.*

***Video proofs,*** *such as this one of a sphere being turned inside out, are transforming mathematics.*

SCIENTIFIC AMERICAN
October 1993 Volume 269 Number 4

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 shows a scene from a computer-generated film of a sphere being turned inside out, or everted. The “video proof,” which was produced at the Geometry Center in Minneapolis, Minn., is based on a topological theorem by William P. Thurston of the Mathematical Sciences Research Institute (see “The Death of Proof,” by John Horgan, page 92). The rules of topology allow the skin of the sphere to be stretched and twisted and even to pass through itself, but the eversion must be completed without the formation of a kink.

## THE ILLUSTRATIONS

Cover illustration by the Geometry Center, University of Minnesota

| Page | Source |
|---|---|
| 32–33 | Miguel L. Fairbanks (*top, first six photographs*), Ken Biggs/Tony Stone Images (*top, last photograph*), Edwin Maynard/The Environmental Picture Library (*bottom*) |
| 34–35 | Jana Brenning |
| 36 | Johnny Johnson |
| 38 | UPI/Bettmann Newsphotos |
| 39 | Alon Reininger/Contact Press Images |
| 42–43 | Ian Worpole |
| 44 | Ocean Drilling Program (*left*), Patricia J. Wynne (*right*) |
| 45 | Patricia J. Wynne (*left*), Australian Geological Survey Organization (*top right*), Millard F. Coffin (*bottom right*) |
| 46–48 | Ian Worpole |
| 49 | Roger Ressmeyer/Starlight |
| 50–51 | Visual Logic |
| 52–58 | Jared Schneidman Design |
| 59 | Thomas C. Halsey |
| 60–62 | Jared Schneidman Design |
| 63 | Thomas C. Halsey |
| 64 | © 1993 Universal Studios/Amblin Productions; Courtesy of Industrial Light & Magic |
| 68–71 | Patricia J. Wynne |
| 72–73 | Paul Alan Cox (*photographs*), Patricia J. Wynne (*drawings*) |
| 74 | Paul Alan Cox |
| 76–77 | Courtesy of Tim Shallice |
| 78 | Johnny Johnson (*top*), Boris Starosta (*bottom*) |
| 79–82 | Boris Starosta |
| 84–85 | Courtesy of Vasa Museum, Stockholm |
| 86–87 | Jana Brenning (*top*), Vasa Museum (*bottom*) |
| 88 | Hank Iken |
| 89–91 | Courtesy of Hans Hammarskiöld, Vasa Museum |
| 92–93 | Geometry Center (*computer art*), Stephanie Rausser (*photograph*) |
| 94 | Robert Prochnow |
| 95 | Per Breiehagen/Black Star (*photograph*); Jean E. Taylor, Rutgers University (*computer art*) |
| 98 | James T. Hoffman, Edward C. Thayer, G.A.N.G., Mathematics Department, University of Massachusetts at Amherst (*computer art*); Jessica Boyatt (*photograph*) |
| 99 | David Ben-Zvi, Geometry Center (*computer art*); Robert Prochnow (*photograph*) |
| 100 | Johnny Johnson |
| 112 | Michael Goodman |
| 113 | Jason Küffer |
| 114 | *Cliff Dwellers* (1913), by George Wesley Bellows (American, 1882–1925); Los Angeles County Museum of Art, Los Angeles County Funds |


# SCIENTIFIC AMERICAN®

*Established 1845*

# LETTERS TO THE EDITORS

## Simplify, Simplify

Essayist W. Brian Arthur ["Why Do Things Become More Complex?" SCIENTIFIC AMERICAN, May] observes that fundamental designs tend to accrete increasingly complex hierarchies of support subsystems as their uses expand, until they are supplanted by improved designs. Arthur then wonders whether a general principle underlies this tendency.

Yes, a well-known characteristic of invention: it is much easier to elaborate than to innovate.

DAVIDSON CORRY
Seattle, Wash.

Arthur writes that "Copernicus's dazzlingly simple astronomical system, based on a heliocentric universe, replaced the hopelessly complicated Ptolemaic system." Copernicus's heliocentric system was actually more complex. Because Copernicus used circular orbits to describe the motions of the planets (as did the Ptolemaic system), he was forced to use epicycles to account for their apparent retrograde motions as viewed from the earth. Not until Kepler discovered that the planets move in elliptical orbits around the sun, some 50 years later, did the heliocentric system emerge in its true simplicity.

DERICK W. OVENALL
Wilmington, Del.

## Vital Priorities

In "The Economics of Life and Death" [SCIENTIFIC AMERICAN, May], Amartya Sen postulates that the problem is not only production but distribution of food. Unfortunately, he does not touch on the political conditions necessary for that distribution. The only time he refers to this issue is when he writes that, in general, "democratic" countries such as India have been able to avoid famine better than "dictatorships" like China.

The available information, however, does not support that statement. Contrary to what Sen indicates, China has been able to feed its people better than India has. Although China had worse health and nutritional levels than India in the 1940s and 1950s, China has been able to improve its health and nutritional indicators faster, at least until the 1980s, when China started the market reforms that increased food production while worsening its distribution. Similarly, Cuba has been able to reduce famine and malnutrition much faster than other Latin American countries, including Costa Rica and Jamaica, the two Sen cited as successes.

The root of the problem of malnutrition in the world is the skewed concentration of wealth that is impermeable to democratic pressures. Only those countries that have redistributed their wealth have been able to solve the malnutrition problem.

VICENTE NAVARRO
Departments of Health Policy,
Sociology and Policy Studies
Johns Hopkins University

We wish to take exception to the article by Sen. The author draws the inaccurate conclusion that life expectancy in Saudi Arabia is lower than in many poorer countries because of mismanagement of health care funds. You must remember that major progress toward modernization in the Kingdom did not begin until 1970. India, in contrast, has a long history of modernity, including British colonization. Thus, Saudi Arabia has lacked the necessary infrastructure required to achieve the levels attained by others.

ABDUL-MANNAN TURJOMAN
Scientific Counselor
Royal Embassy of Saudi Arabia
Cultural Mission to the U.S.A.
Washington, D.C.

***Sen replies:***

Navarro is right to argue that unequal distributions of wealth and income have much to do with undernourishment. The need for public services, on which I focused, relates to this inequality.

His argument about India and China confuses famine with endemic undernourishment and ill health. They are distinct problems. I had, in fact, noted that "even though postrevolutionary China has been much more successful than India in economic expansion and in health care, it has not been able to stave off famine." Navarro overlooks the gigantic famine of 1958–1961 in China, which is estimated to have killed from 23 million to 30 million people. India has not had a famine since its independence. Any government that had to face opposition parties, free newspapers and regular elections could not afford to have famines. During those years of famine, the Chinese government was under little political pressure to abandon the policies that were directly contributing to the problem. A similar governmental immunity sustained famines in Cambodia in the 1970s, in the Soviet Union in the early 1930s and in many nondemocratic regimes—both of the "right" and the "left"—in sub-Saharan Africa. More democratic regimes in Botswana and Zimbabwe successfully prevented famine.

I did not attribute the relatively low life expectancy in Saudi Arabia to "mismanagement of health care funds." Rather the main issue concerns the overall priority that is given to health care and education (especially female education) compared with other expenditures. To attribute the achievements of the Indian state of Kerala to the benign influence of the British Empire would be mistaken on two grounds. First, most of Kerala remained outside British India. Second, the biggest expansions of education and health care in Kerala have come only in recent decades, much influenced by left-wing political movements.

## What It Was, Wasn't

I admire presidential science adviser John H. Gibbons as much as the next person ["Profile," "Science and the Citizen," by John Horgan; SCIENTIFIC AMERICAN, April], but I fervently hope—for the sake of the nation—that his expertise in technology is superior to his knowledge of popular culture. Gibbons quotes a supposed Tennessee Ernie Ford story about a mountaineer watching a football game. Alas, it wasn't Tennessee Ernie Ford, it was Andy Griffith in the 1953 comedy routine that made him famous, *What It Was, Was Football*. Poor Gibbons mangled other details, too (although, in Griffith's original, the tale spinner *was* drinking a big orange drink).

HOWARD R. COHEN
Los Angeles, Calif.

*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

OCTOBER 1943

"For many years it was believed that the isolation of a protein in the chemically pure state was nearly impossible. Recently, however, this situation has changed quite completely. The enzymes pepsin, trypsin, and urease and the hormone insulin have all been isolated and appear to be proteins. One of the brightest chapters in this search for protein in its chemically pure state is the isolation of the virus of the tobacco mosaic disease. This virus is described as a crystalline protein. The significance of this discovery lies not only in the great advance in the understanding of protein structures but in the fact that it connects proteins with measles, yellow fever, the common cold, and several other diseases of both plants and animals."

"Motors running on alternating current have only a few fixed speeds depending upon the motor construction and the frequency of the current supplied to them. What industry desires is the flexibility of the direct-current motor and the efficiency of transmission of the alternating-current system. Electronics now makes this possible. The tubes employed are thyratrons—gaseous or mercury-vapor-filled rectifiers which convert alternating current into direct current. The portion of the positive half cycle during which current flows through the tube can be varied by merely changing the voltages on the grid with respect to the phase of the alternating voltages placed upon grid and anode."

"When an optical surface of glass is aluminized, the evaporated molecules, being in a high vacuum, travel, without bumping into other molecules, from the hot metal source to the mirror's cold surface and are deposited in a noncrystalline metallic film having the same degree of polish as that of the glass. As soon as air is admitted, the metallic aluminum begins to oxidize, and this oxide continues to thicken for about 60 days. Why couldn't all this be as easily accomplished by letting a disk of plain cast aluminum oxidize in the air? Fred B. Ferson, a Biloxi, Mississippi, amateur telescope maker, states it thus: 'Aluminum is a metal which absorbs gases readily, and is hard to prevent from taking up impurities when it is cast. Also in castings it cools into crystalline structure, the crystals coarse and full of holes—possibly from absorbed gases driven off.'"

OCTOBER 1893

"It now does not seem improbable that, when by the power of thought an image is evoked, a distant reflex action, no matter how weak, is exerted upon certain ends of the visual nerves, and, therefore, upon the retina. Helmholtz has shown that the fundi of the eyes are themselves luminous, and he was able to *see,* in total darkness, the movement of his arm by the light of his own eyes. This is one of the most remarkable experiments recorded in the history of science, and probably only a few men could satisfactorily repeat it, for it is very likely that the luminosity of the eyes is associated with uncommon activity of the brain and great imaginative power. It is fluorescence of brain action, as it were.—*Nikola Tesla, in a paper read before the Franklin Institute.*"

"Japanese children are suckled until their sixth year, and in language unmistakable may be heard asking for the lactatious fountain. In view of the almost universal use of cow's milk in other countries, its exclusion from the diet of the Japanese raises the interesting subject of inquiry as to whether or not the race benefits by this custom, and Dr. A. S. Ashmead, of New York, discusses the question in the *Sei-i-Kwai* medical journal. In the first place it is assumed that indirectly the absence of cow's milk is most beneficial. The Japanese mother feels the compulsion of looking after her own health and diet. Japanese mothers chiefly live on rice, 'fish, shells, seaweed, and other products of the sea,' while wine and beer are rigidly excluded. The reward of all this meritorious care of motherhood and childhood is the absolute freedom from rickets. Again, the author holds that the transmission of tuberculosis is avoided by the exclusion of cow's milk from the infant's dietary."

"Efforts have been made to teach a child how to swim by supporting him in the water and causing him to effect the motions of natation. This is the most practical process. Its inconvenience is that it necessitates the presence of a teacher with each pupil, and, in a large class of children, the teacher cannot occupy himself with each of them for a very long time. Mr. Devot has been able to overcome all the difficulties of the preceding method in a very ingenious manner. His apparatus (*below*) permits the pupil to learn in conditions entirely identical with those that present themselves when he tries to sustain himself alone in the water. The apparatus is in use among the pupils of the Michelet Lyceum, who have been the first to benefit from the invention of their master, Mr. Devot.—*La Nature*."

***Apparatus for teaching swimming***

## Hard Times

*Occupational injuries among children are increasing*

MARK PETERSON SABA

***SWEATSHOPS in New York City's garment district have been found to employ young people illegally. Many such factories have dangerous working conditions.***

The sepia photographs and etchings lie in the archives of nearly every industrialized nation: children in factories operating or dodging dangerous machinery, working too many hours for too little pay. Now, for the first time since U.S. labor laws protecting children were enacted in 1938, such stark images are being recorded again in record numbers. The number of children working—legally and illegally—has risen dramatically throughout the country in the past decade. Accompanying this increase are numerous, sometimes grisly, reports of labor law violations as well as occupational injuries—amputations, electrocutions, fractures, burns, lacerations—and deaths.

Federal law prohibits anyone under the age of 14 years from working, except in some agricultural jobs, and anyone younger than 18 years from performing certain tasks, such as operating heavy machinery or working on construction sites or with toxins. Nevertheless, nearly every industry employs children of all ages. About five million teenagers work in the U.S. In 1991 an additional 27,528 children were discovered in illegal jobs, an increase of about 300 percent since 1983, according to the General Accounting Office (GAO). The number of youths working clandestinely has been estimated to be as high as 676,000.

Few studies chronicle the hazards experienced by these young people. The U.S. Department of Labor keeps comprehensive records only of wage and hour infractions: children working for less than minimum wage or for excessive hours. By this measure alone, the country is not doing well. In 1990, the year of a major federal effort called Operation Child Watch, the department uncovered 39,000 such violations. In 1991 and 1992, years when no special operation was in place, about 29,000 and 19,000 violations were reported, respectively. Advocacy groups argue that the figures are actually much higher. "We have no compunction about saying that there are over two million child labor violations each year," asserts Jeffrey Newman of the National Child Labor Committee in New York City.

Estimating the number of injuries among these children is even more difficult. The statistics that are available rely on worker's compensation data and hospital records that are, at best, incomplete. According to child labor experts and pediatricians, children working for family businesses often do not report injuries to worker's compensation boards. In most states, agricultural labor is not even covered by worker's compensation.

Other youths, particularly those employed illegally, can be unaware of the procedure or can be discouraged by their employer from making a claim. A 1988 survey by the New York State Department of Labor, for instance, found that more than half of the teenagers questioned had experienced an injury or a wage or hour violation. Only one third of those injured reported the accident. Underreporting may also occur because physicians generally do not record information related to a young patient's occupation, notes Philip J. Landrigan of the division of environmental and occupational medicine at Mount Sinai Medical Center in New York City.

Despite gaps in the data, the National Institute for Occupational Safety and Health determined that more than 130 children died on the job in 1990. Several other studies have looked at state cases. Landrigan and his colleagues recently reported in the *Journal of the American Medical Association* that 9,656 New York State adolescents received worker's compensation between 1980 and 1987. Of those, 43.5 percent experienced a permanent disability—that is an average of 525 teenagers a year—and 31 children died.

In a study to be published in the *American Journal of Industrial Medicine,* the Massachusetts Department of Public Health found that the rate of occupational injury among teenagers was about twice that of adults. Of the 17-year-olds whose location of injury was recorded when they were treated in sampled emergency rooms, 26 percent were there because of a work-related injury. And researchers at the University of Washington in Seattle determined that farm work accounted for nearly 50 percent of the serious injury claims among children 13 years old and younger. According to an article in the *American Journal of Public Health,* children aged 19 years and younger are involved in 25,000 farm accidents around the country annually; 300 are killed.

The reasons for the increases in child labor are varied. Landrigan and others cite poverty and immigration as contributing factors. Indeed, according to the Children's Defense Fund, there are 14.3 million poor children in the U.S., the highest number since 1965. Many of them have difficulty finding aboveboard jobs, notes Joseph A. Kinney of the National Safe Workplace Institute in Chicago. The GAO has found that minority and low-income children are more likely to be employed in hazardous occupations than are their white or higher-income counterparts.

In New York City, for example, some 1,500 sweatshops in the garment industry are a large source of illegal work— and child labor violations. "Most of the children in these factories are below 18, some as young as eight," notes Thomas Glubiak, chief of the city's 20-member garment district task force. These youths are exposed to myriad dangers, he notes. "We find unguarded machinery, no fire exits, boilers, wiring problems, egress problems, machines too close together." According to the New York State Department of Labor, the number of city establishments illegally employing children rose from 19 in 1987 to 122 in 1988.

At the same time as more children have entered the work force, advocates argue that enforcement of child labor laws has diminished. Many experts consider the labor laws to be strong, indeed, most states have legislation that is more stringent in certain aspects than federal law. But they perceive follow-up to be weak. "Active enforcement is absolutely essential," Newman notes. "This is one of the areas where you can actually stop the crime from happening."

Budget cuts and inadequate staffing have limited the ability of the federal Department of Labor to conduct sweeps, Newman says. The GAO has reported that the wage and hour inspectors follow up on repeat offenders infrequently. And a National Safe Workplace Institute report stated that only 11 percent of the inspectors' time is spent on child labor; "an establishment that employs adolescents can anticipate a federal inspection once every 50 years," it estimates. The Department of Labor counters that its 850 inspectors check for child labor violations constantly. Regardless of the department's claim, however, two bills before Congress cite enforcement as the primary deficiency in U.S. law and seek to increase such efforts and associated penalties.

Internationally, conditions look more dire. According to the International Labor Organization (ILO) in Geneva, some 200 million children under the age of 15 years work. The numbers "seem to be increasing in the Third World, where you have worsening economic conditions and worsening political situations," notes Susan E. Gunn of the ILO. In Brazil, 18 percent of children aged 10 to 14 years work—a total of seven million. "But frankly, it is very hard to generalize from country to country," Gunn says. For instance, in Thailand the number of working children rose 34 percent to an estimated 1.7 million between 1983 and 1987 while the economy boomed: exports rose 84 percent.

If it is difficult to find figures on the extent of occupational injuries in the U.S., it is virtually impossible worldwide. Isolated reports hint at the extent of the problem. For instance, this spring a widely publicized factory fire near Bangkok left 500 people injured and 188, mostly young women, dead. The workers were making dolls for Kader Industrial, a supplier for the manufacturer of Cabbage Patch Kids. Whereas in the U.S. more young boys are injured, in other parts of the world girls are more likely to be abused. Employers "favor hiring girls—they feel girls are more exploitable," says Robert Senser of the Asian-American Free Labor Institute in Washington, D.C. "Most of the workers in the garment sectors are girls. They are considered more docile and more dexterous."

The recent attention turned on youth employment is not limited to the U.S.— international concern is rising as well. The ILO reports that it is stepping up its programs. Although there is no mechanism for enforcement, Gunn notes that moral pressure can be effective. In addition, American consumer groups and two congressional bills are seeking to halt imports of products that have been made by children under 15 years of age. The Indian government has already responded to the threat of sanctions with promises of more crackdowns on child labor violators and a system of labeling to ensure products were made without exploiting children. According to the ILO, about five million Indian children work as bonded laborers.

But, as with all aspects of this problem, the impact of trade barriers on the lives of these children has not been studied. "I do not have an opinion yet," Gunn says. "It is obviously having an effect, but positive or negative? We just don't know." —*Marguerite Holloway*

## Sentries and Saboteurs

*Mutating patients' genomes to suit their medicine*

Doctors treating cancer with chemotherapy play the odds. Too little drug and the diseased tissue will survive; too much and the damage to healthy cells will be more than the patient can bear. But some researchers believe gene therapy can increase the spread. In upcoming clinical trials they intend to insert into cancer patients' DNA mutations that either make healthy cells more tolerant of existing medication or make tumors more vulnerable to it. "This is a totally novel approach, but a natural extension of gene therapy," says Albert B. Deisseroth of the University of Texas M.D. Anderson Cancer Center in Houston.

In June the Recombinant DNA Advisory Committee of the National Institutes of Health approved two protocols for human tests of a strategy that Deisseroth calls "chemoprotection." He and others plan to genetically modify patients' bone marrow cells so they can withstand more courses and higher doses of chemotherapy. Human trials are also under way to test another gene therapy technique that is essentially a mirror image of chemoprotection. Rather than inserting "sentries" to guard healthy cells against dangerous drugs, a strategy called molecular surgery sends in genetic "saboteurs" to make tumor cells susceptible to a normally harmless medicine.

The Texas group headed by Deisseroth plans to enlist 20 to 30 women whose late-stage ovarian cancer has already bested standard surgery and drug treatment. The most effective medicine available for such cases is often Taxol, a toxin derived from the bark of the Pacific yew tree. Taxol kills cells caught in the act of dividing; since tumor cells multiply more rapidly than do normal cells, they are more vulnerable. Unfortunately, Deisseroth adds, "the major side effect of Taxol is that it also kills bone marrow cells," which produce the white blood cells of the immune system. Just two or three courses of Taxol are enough to deplete most patients' marrow so severely that they succumb to infection or internal bleeding.

Two or three courses are often not enough to vanquish the slow-growing tumors of ovarian cancer. So Deisseroth wants to remove 10 percent of each patient's bone marrow before chemotherapy begins. Using an antibody that binds only to younger marrow cells, the researchers will try to separate out the stem, or progenitor, cells that have yet to differentiate into a particular kind of white blood cell. They then plan to infect these stem cells with a mouse virus—crippled so that it cannot reproduce and spread—into which has been inserted the human MDR gene for multiple-drug resistance.

All marrow cells naturally have the MDR gene. When turned on, it produces a protein that acts as a sentry, escorting toxic chemicals such as chemotherapeutic agents out of the cell before they can do any damage. But for some unknown reason the gene is turned off when a stem cell matures into a white blood cell, so only stem cells are protected and then only weakly. "The guard is easily overrun," Deisseroth says. "The higher the concentration of the drug, the more sentries you need to protect each cell."

By infecting stem cells with MDR-bearing viruses, the researchers hope to secure them against Taxol. After each patient completes an ordinary round of chemotherapy, she will be injected with her own mutated stem cells. Many of those cells should carry multiple copies of the MDR gene and an extra copy of

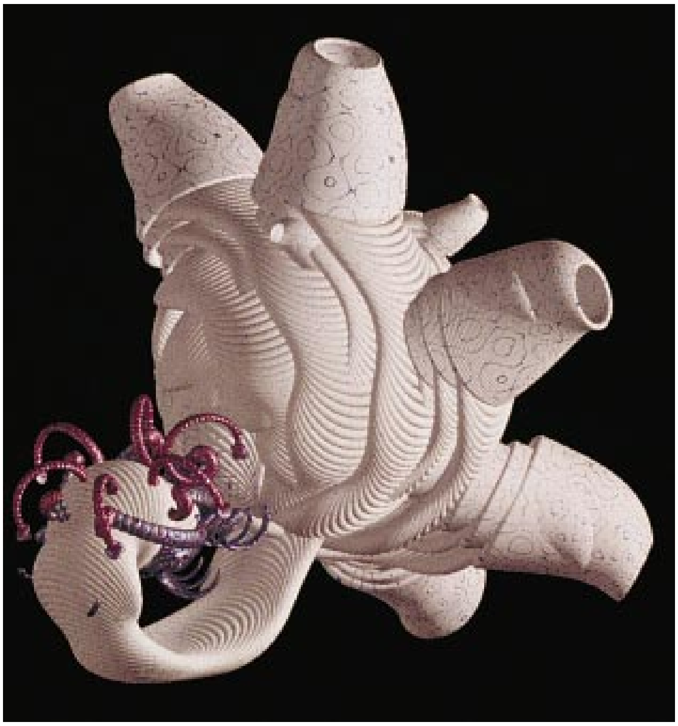
## Creative Evolution

Like all things cultural, art evolves—a platitude to most artists, but to a handful, a description of technique. By appropriating scientific principles from evolutionary biology and advanced technology from computer graphics, this avant-garde has fashioned a new medium for creative expression: virtual evolution.

British artist William Latham designed software to mutate the "genome" of three-dimensional forms. From each generation he selects and breeds the most aesthetically "fit."

Karl Sims, artist-in-residence at Thinking Machines, discovers his artworks using a self-mutating program that draws from a genome of equations and transformations. "You quickly evolve equations that you couldn't design or even understand," he says. Nevertheless, the process is simple enough that museum-goers can express their taste via interactive exhibits.

More dynamically, neural networks add collective behavior to an ecosystem of intelligent paint brushes bred by the genetic algorithms of Michael Tolson, chief scientist of Xaos Tools. Make way for virtual surreality. *—W. Wayt Gibbs*



WILLIAM LATHAM

WILLIAM LATHAM

***SURVIVAL OF THE MOST INTERESTING** is the rule when artists play God and gardener. Beginning with simple objects, William Latham breeds myriad generations of computer-generated mutations to arrive at his pseudo-organic "White Form"* **(left)**. *His creations*

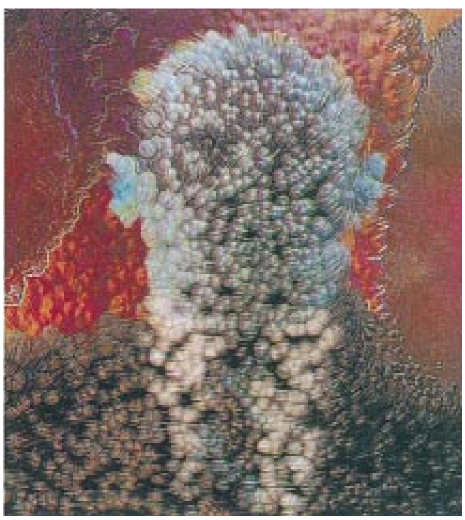
KARL SIMS *Thinking Machines*

KARL SIMS *Thinking Machines*

***SUPERCOMPUTER** and image evolution software installed by Karl Sims at the Pompidou Center in Paris let visitors view his art* **(left)** *or interact with the exhibit to create their own collaborative works* **(above)**.

the regulatory machinery needed to produce it. After a few weeks of recovery time, Deisseroth will begin giving Taxol again, in gradually increasing doses. With each course, more of the unmodified marrow cells—greater than 90 percent of the marrow at first—will die. But if the progeny of the transgenic stem cells keep making the drug-resistance protein, they should prosper. If all goes

*can display characteristic behavior and respond to their surroundings, like the chameleonic "Mutation in Red Room"* **(above).**



MICHAEL TOLSON *Xaos Tools*

***"OUR FOUNDER," part of a series by Michael Tolson awarded the Goldene Nica Award at the 1993 Prix Ars Electronica in Austria.***

# Never Give a Sucker an Even Break

Shall the meek inherit the earth, or is might right? Students of behavior have expended much effort analyzing whether evolution should favor individuals who cooperate or exploiters who go for short-term gains. A computer tournament over a decade ago indicated that a cooperative behavior strategy known as tit-for-tat can beat out exploiters, although it is not stable in more lifelike simulations. Recent computer modeling has uncovered a more resilient formula, dubbed Pavlov, that prevails over tit-for-tat in the evolutionary race and yet is still cooperative—up to a point. Its discoverers believe Pavlov could be a model for many examples of altruism in the natural world.

The standard paradigm for analyzing the evolution of cooperation is a conundrum called the Prisoner's Dilemma. This problem considers two suspects, imprisoned separately, each contemplating whether to confess and plead for a light sentence (to cooperate) or to blame the other for the crime (to defect). The payoff for each player depends on what the other does. If both cooperate, they each do better than if they both defect or exploit the other. Yet each could do better by defecting than by cooperating [*see table*].

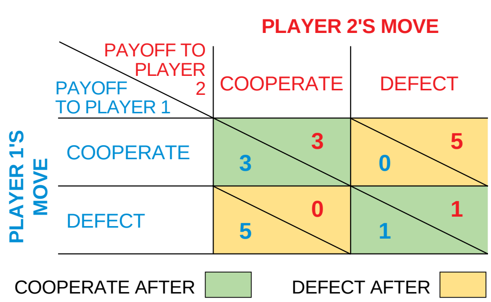
Although defecting always pays off in a single round, cooperative strategies can be advantageous if the Prisoner's Dilemma is repeated—as interactions between members of a community often are. When Robert Axelrod, a professor of political science at the University of Michigan, organized a contest of computer programs to play the Prisoner's Dilemma in the early 1980s, tit-for-tat was the surprise winner. It was a surprise because it was so simple: tit-for-tat starts by cooperating and in subsequent bouts simply repeats what its opponent did in the previous round.

Martin Nowak of the University of Oxford and Karl Sigmund of Vienna University showed two years ago that tit-for-tat does not endure in simulations in which, as in the real world, occasional mistakes are made. Eventually, they found, it gets bogged down in bouts of backbiting. Recently Nowak and Sigmund have conducted simulations that also model the effect of mutations, or small changes in strategies. The results, published in *Nature*, were every bit as startling as those from Axelrod's tournaments.

Under a wide range of conditions Nowak and Sigmund found that if they waited long enough, the strategy they call Pavlov usually dominated their modeled populations. A player following Pavlov cooperates after a contest in which both parties cooperate—and also after a contest in which both parties defect. If only one player cooperated on the previous round, however, Pavlov defects. The strategy is so named because, put another way, it repeats, reflexlike, its previous play if it gets one of the two higher payoffs and switches if it comes out behind.

A population of Pavlovian individuals does cooperate and reap the attendant benefits. But a Pavlovian population has no tendency to start cooperating indiscriminately. The weakness of tit-for-tat, Nowak and Sigmund say, is that mutation allows populations to become more and more cooperative, which eventually leads to an invasion by selfish "always-exploiters."

Nowak and Sigmund note that animals often retaliate against defection. These findings have been taken as support for the idea that animals follow a tit-for-tat strategy. But such results also corroborate the idea that animals use a Pavlov-like strategy. Pavlov, while cooperative with others who cooperate, has "no qualms about exploiting a sucker," according to its discoverers—it keeps on defecting against an opponent foolish enough to go on cooperating. And that ruthlessness is what allows Pavlov, and cooperation, to survive. —*Tim Beardsley*




***THE PRISONER'S DILEMMA is illustrated by the benefit (numbers of points) each player gains depending on whether he or she cooperates or defects. In sophisticated simulations, the strategy called Pavlov* (shown in green and yellow) *usually eventually predominates.***

well, the patient's entire marrow and blood supply will grow to tolerate Taxol while the tumors wither away.

Deisseroth admits that much of the process he hopes to modify remains unknown. The virus inserts MDR and its replication sequence randomly into the stem cell genome; the researchers rely on luck to avoid causing unwanted mutations. No one knows why MDR is naturally turned off in mature white blood cells. There may be a good reason, and short-circuiting that "off" switch may produce side effects.

Since ovarian cancer does not spread to the bone marrow, there appears to be little danger of making tumors themselves drug resistant. But as a patient's cumulative dose of Taxol builds, it will take a toll on other tissues. Raising the typical Taxol dosage just 40 percent has been shown to cause neurological problems, says Charles Hesdorffer of Columbia University, whereas other chemotherapeutic drugs can be increased five- to 10-fold without harm. Hesdorffer sponsored the second, recently approved chemoprotection protocol, which will explore other drugs in addition to Taxol and will also include patients with widespread breast cancer or brain tumors.

Deisseroth is confident nonetheless. "We've tried this out in mice, and it works," he claims. "As far as we can tell, the cells are not changed by the MDR gene except in their ability to repel toxins." But Arthur W. Nienhuis, whose group at the NIH has collected much of the animal data cited by Deisseroth and has submitted a third chemoprotection protocol, warns that mice may not be the best model. "In primates—and it seems to be true in humans as well—only 1 to 5 percent of the modified stem cells actually produce MDR, versus 20 to 50 percent in mice," he says. That may be too little too late. "I think we need to make sure that we're getting the gene in, that it's safe and that it produces protein at a reasonable level," Hesdorffer adds. "Then we can go on to determine whether this system actually works therapeutically."

Cautious optimism also characterizes experiments to test the molecular surgery technique spearheaded by Kenneth W. Culver of Iowa Methodist Medical Center, Edward H. Oldfield of the NIH and Genetic Therapy in Gaithersburg, Md. In experiments last year Culver and Oldfield managed to rescue 11 of 14 rats with terminal brain cancer. The first human trial began last December at the NIH. Culver recently won approval for two more trials to begin in November at Iowa Methodist and at Childrens Hospital in Los Angeles.

Molecular surgery sidesteps some of

the special difficulties brain cancer presents and capitalizes on the unique opportunities it offers. Brain tumors often lurk too deep for scalpels to reach, and the blood-brain barrier can filter out drugs before they reach their target. But once a brain has matured, its cells largely stop dividing, so anything that multiplies is probably malignant.

Culver came up with the idea of sneaking a gene from the herpes simplex virus into tumor cells and then killing them with the antiherpes drug ganciclovir. The gene, which codes for the enzyme thymidine kinase and is thus known as HS-tk, can be carried into a cell by a retrovirus only when the cell is dividing, so normal brain cells should be safe. But to get enough of the engineered virus into the brain to do any good, Culver had to genetically transform mouse cells into viral factories.

Culver and Oldfield found that when they injected these virus-producing cells into brain tumors in rats, as much as 60 percent of the tumor cells incorporated the HS-tk gene, whereas virtually no normal cells did. When they then gave the rats ganciclovir, they got another pleasant surprise. The drug killed not only the infected tumor cells but also many malignant cells that did not carry HS-tk. In some cases, even when only 10 percent of the cancerous cells took up the gene, the whole tumor disappeared without causing inflammation or hemorrhage.

This bizarre "bystander effect" does not seem to extend to normal cells surrounding tumors. "We still don't know how this works," Culver admits. "One hypothesis is that it involves intercellular communication." Destruction of the circulatory system within the tumor may also play an important role, effectively starving to death those tumor cells that do survive the drug.

Human trials have yet to prove that molecular surgery is as precise and thorough in people as it seems to be in rats. And, Culver points out, this technique is not without its problems and risks. The virus-producing cells must be injected close to the cancer for the treatment to work, but magnetic resonance imaging does not always reveal every tumor. If the mouse cells (which also contain HS-tk and so also die when exposed to ganciclovir) manage to escape, they could infect other parts of the body, such as the bone marrow.

For patients who have few options and little hope, however, molecular surgery and chemoprotection may prolong—perhaps preserve—life. If so, this twist on gene therapy could give doctors a new weapon in the frustrating battle against cancer. —*W. Wayt Gibbs*

## Sharks Do Get Cancer

*Cartilage cure relies on wishful thinking*

A pseudoscientific myth holds that sharks don't get cancer. Indeed, that proposition is the title of a book by I. William Lane and Linda Comac promoting shark cartilage as a "breakthrough in the prevention and treatment of cancer and other degenerative diseases." The CBS television program "60 Minutes" enthusiastically picked up the idea earlier this year.

Unfortunately, John C. Harshbarger, director of the registry of tumors in lower animals at the Smithsonian Institution in Washington, D.C., says he has records of at least 20 cases of cancer in sharks. The registry, which is supported by the National Cancer Institute, includes shark cancers that originated in cartilage, as well as cancers of the kidneys, liver and blood cells. (Harshbarger says data do not exist to determine whether sharks get cancer more or less often than do other creatures.)

In their book Lane and Comac cite an article about cartilage by Arnold I. Caplan, a biologist at Case Western Reserve University, published in this magazine in October 1984. Caplan had noted that cartilage seemed to contain substances that inhibit the growth of blood vessels and speculated that such substances might someday be used to prevent tumors from establishing the blood supply they need to grow.

Caplan was right: cartilage does indeed contain inhibitors of blood vessel growth, or antiangiogenesis factors. Some have been isolated and are now in clinical trials for treatment of a variety of cancers as well as stomach ulcers. But Caplan says he is "appalled" by Lane and Comac's effort to promote shark cartilage as a treatment. "This is an extreme interpretation of the data," Caplan complains. Moreover, he adds, "There's nothing special about sharks. Why not eat pigs' knuckles?"

Caplan is not the only investigator troubled by the shark cartilage craze. Judah Folkman, an angiogenesis researcher at Harvard Medical School, states firmly that there is no evidence from any controlled study that ingesting cartilage can treat cancer. What is more, he notes, on the basis of what is known, "a patient would have to eat hundreds of pounds of cartilage" to have any chance of experiencing an effect. Folkman says he has been trying for years to dissuade Lane from using his name to promote a shark cartilage product.

Carl A. Luer, an investigator at the Mote Marine Laboratory in Sarasota, Fla., who appeared on "60 Minutes," is also dismayed that his work on antiangiogenesis factors is being used to promote shark cartilage. "I feel our factor is a protein and can't be absorbed," he states. Some tropical shark populations, he says, are already depleted by overfishing.

The National Cancer Institute has examined some cases of supposed improvements in patients treated with shark cartilage, but the data "did not show solid evidence of clinical activity," according to Mary S. McCabe, a clinical trial organizer at the institute. Still, enthusiasts are persevering. Charles Simone, a physician in private practice in New Jersey, maintains that a few of his patients have responded favorably. But even he warns against self-medication. Much of the shark cartilage now on sale in health food stores, he says, is "bogus stuff." —*Tim Beardsley*

## Run Silent, Run (Not So) Cheap

For decades, polar oceanographers have wanted to come in from the cold. To do their work, they have had to lug cumbersome sonar and seismic equipment over the surface of polar ice floes. Meanwhile, a few fathoms below, the crews of Soviet and U.S. submarines have lived in relative comfort as they played cat-and-mouse games while practicing for the end of the world.

Some of those frostbitten scientists may finally get a chance to do their work in the environment enjoyed by their undersea neighbors. This past summer an SSN-650 attack submarine, the *Pargo,* set off from Groton, Conn., with five civilian scientists nestled among the more than 100 regular crew members. The mission, Submarine Arctic Science Cruise-93, which is to be funded with $3 million from this year's defense budget, constitutes the first unclassified scientific journey on board a nuclear submarine. On August 11, this Sturgeon-class sub set out on a journey of more than a month to the Arctic Ocean that would take the researchers to the North Pole. Almost half the mission, 3,300 nautical miles, was to be below the ice.

The "hunter-killer" submarine is most likely still equipped with a cargo of Harpoon antiship missiles and Mk 48 torpedoes in the unlikely event that hostilities break out. But this expedition could mark a step toward perhaps the ultimate defense-conversion project: an attempt to deploy a nuclear submarine for the sole purpose of scientific research.

Since the first nuclear-powered submarine was launched in the 1950s, scientists have recognized that these vessels might be harnessed to compile a comprehensive profile of physical changes in the ice cap, an especially important task in tracking the possible warming trend in the earth's climate. "It would be difficult to get this information any other way," says Marcus G. Langseth, a scientist at Columbia University's Lamont-Doherty Earth Observatory and chairman of the science steering committee for this program.

The current mission was to use a type of sonar that can look upward to judge the volume of the ice cap. Meanwhile a synthetic-aperture radar on a satellite orbiting over the Arctic would provide topside images of the same area. The submarine also plans to surface in holes or cracks in the ice. These stops would allow researchers to deploy buoys with strings of sensors that hang under the ice to measure water temperatures and salinity.

Langseth, however, believes that researchers will be able to take full advantage of a submarine only once it is permanently retrofitted with the necessary instrumentation. A high-level study panel of oceanographers recommended to the navy in 1991 that magnetic and seismic instruments for charting the Arctic basin could be towed behind the vessel. For the moment, this technology would probably be unacceptable to the navy: submarine commanders assiduously avoid any devices that emit substantial amounts of radio energy for fear of revealing submarine location and operating characteristics.

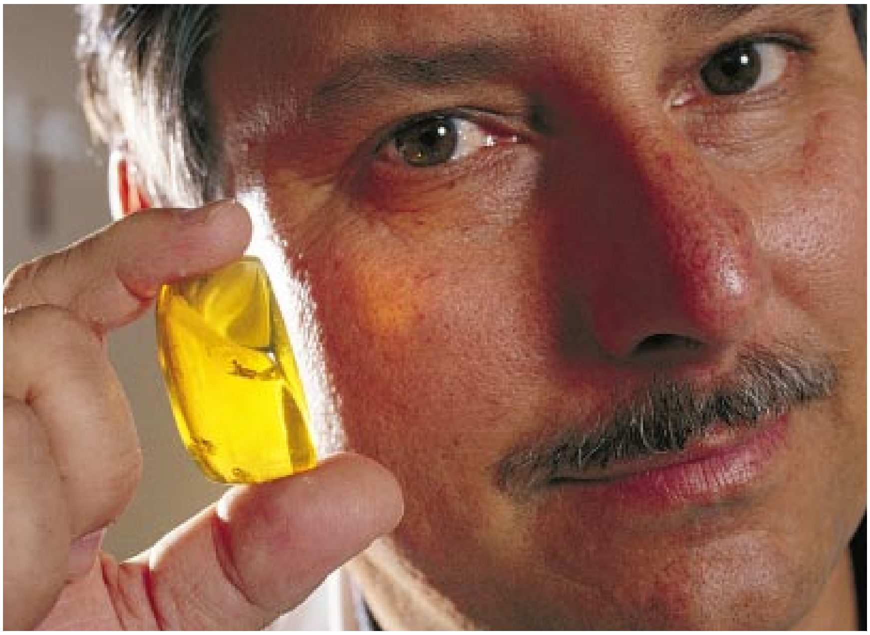
So the oceanographers could run silent. But can they run cheap? A decision about a dedicated research vessel, Langseth says, should come before the planned decommissioning of the entire Sturgeon fleet. Langseth and other scientists have their work cut out: the $8 million to $15 million needed to keep a submarine in operation annually could make a submarine into the oceanographic equivalent of the Superconducting Super Collider. "The big problem is that the navy doesn't want to pay for it, and the scientific community doesn't want to pay," remarks George B. Newton, a former submarine commander who helped to organize the mission.

Are the Russians coming? Our former adversaries may ignite a fare war for scientist passengers. They have twice offered to convert one of their nuclear submarines into a research vehicle for the West.

Hitching a ride on a Russian submarine, even as part of a U.S. foreign aid package, may prove a hard sell at a time when U.S. defense companies are laying off workers. Indeed, there may be a way to get at least some of the information with little cost. If Congress ultimately decides that a multimillion-dollar passage on board a submarine is an unneeded joyride, the U.S. research community might pressure the navy to speed up the declassification of Arctic data that they have been gathering for years. —*Gary Stix*

HARRY GERWIEN *Courtesy of U.S. Navy*

***STURGEON-CLASS SUB, like this one, was to have surfaced in the Arctic with five scientists on board.***



STEVE STARR SABA

*BUGS INSIDE BUGS: Raúl J. Cano and an amber-clad, bacteria-laden insect.*

## Jurassic Virus?

*Can't clone a Tyrannosaur? Then try chicken pox*

Molecular archaeologists immediately cried foul when they saw the visions of dinosaurs cloned from the last meals of amber-entombed mosquitoes in the film version of Michael Crichton's *Jurassic Park*. Ancient DNA is too damaged, they argued; it would take centuries to piece it together. Moreover, the reconstructed genome would be riddled with errors. In short, they concluded, extinct organisms must remain so, forever.

Not necessarily, say George O. Poinar, Jr., and Raúl J. Cano, the entomologist-microbiologist team who hold the record for sequencing the most ancient DNA—that of a Jurassic wood beetle embedded in amber more than 120 million years ago. They believe it might be possible to resurrect organisms, albeit ones much simpler than *Tyrannosaurus rex*—a bacterium, say, or a virus. "We're very close to reconstructing a gene of a bacillus species that lived in the gut of a stingless bee 40 million years ago," says Cano, who is at California Polytechnic State University in San Luis Obispo. "You'd put the genes into a living species bit by bit, until nothing was left of the modern DNA."

The bacterial gene, known as 16S rRNA, helps make ribosomes. So far, the researchers say, they have strung together 1,300 of a possible 1,500 or so nucleotide bases. They concede that some genes are harder to reconstruct but contend that the entire genome of a bacterium could be built in a few years.

Yet bringing an ancient bacterium back to life might not be so simple. Michael A. Goldman of San Francisco State University points out that bacterial DNA forms loops having quasichromosomal status—that is, it encodes information that lies outside the sequence of nucleic acid bases. In addition, bacteria bequeath proteins as well as nucleic acids to their daughter cells, and these proteins may influence the expression of certain genes. "Already that makes the task different from just making DNA and throwing it into a host cell's genome," he argues.

But Poinar and Cano retort that they will simply shoot at a smaller target. "We hope to be looking into different kinds of viruses pretty soon," says Poinar, who is at the University of California at Berkeley. The smallest ones have genomes no larger than a fair-sized gene and can be made to replicate without helper proteins. One merely inserts the naked viral DNA into a host cell, which then produces complete viral particles. Experts in viral evolution generally agree that this procedure can be done. "Jurassic virus? Sure," says Stephen S. Morse of the Rockefeller University.

The problem, of course, will be isolating an ancient virus. No virus fossils are known, although the new techniques of enzymatic amplification may yet wring viral genes from ancient remains. "Theoretically, resurrecting a virus would be possible, but it may be hard to get good enough DNA," says Peter M. Palese of Mount Sinai Medical Center in New York City. Palese and others have been seeking a more elusive quarry: the virus that caused the massive 1918 influenza pandemic. Such work is particularly challenging because the genes are encoded in RNA, which is less stable than DNA.

Virologists want to know more about the 1918 strain because another such pandemic can, in principle, recur. Until recently, workers had to proceed indirectly, inferring the viral structure from antibodies produced by survivors of the infection or extrapolating from the genomes of more recent viral samples. Walter M. Fitch of the University of California at Irvine has done so by charting the evolutionary trajectories of influenza in humans and pigs. He found that the strains indeed diverged from a common ancestor around 1918.

Although Fitch believes he has reconstructed the genome of the common ancestor, he says it requires correction by direct sequencing of the 1918 virus. While Palese says his attempts to isolate the virus from human remains have been unsuccessful, Robert G. Webster of St. Jude Children's Research Hospital in Memphis says he is sequencing the 1918 virus. He refuses to discuss the results, however.

Another logical RNA target is the HIV retrovirus, which causes AIDS. "You can go into museum collections and select human tissues preserved in the 19th century, when they used spirit rather than formaldehyde," Morse says. "We know that some diseases killed quickly in those days. Galloping consumption was not unlike the presentation of tuberculosis in AIDS patients today. Maybe these were individuals with AIDS who died of secondary infection."

So far, however, the only viral RNA to come from old human remains are endogenous retroviral genes, so called because they became integrated in human chromosomes long before *Homo sapiens* evolved. In a paper that is to appear in *Biological Anthropology and the Study of Ancient Egypt,* Jaap Goudsmit and his colleagues at the Academic Medical Center in Amsterdam report that an Egyptian mummy more than 5,000 years old has yielded traces of human endogenous retrovirus type C. More recent studies on mummified monkeys produced similar results.

Microbiologists do not seem overly worried about disturbing the sleep of undead germs. Most microbes are innocuous, and most laboratories are designed to contain pathogens. But the risks are not zero. "What was a minor microbe in the time of the mummies might be lethal today," Goldman notes. "I would keep in mind Michael Crichton's earlier novel, *The Andromeda Strain*."—*Philip E. Ross*

# PROFILE: FREDERICK SANGER

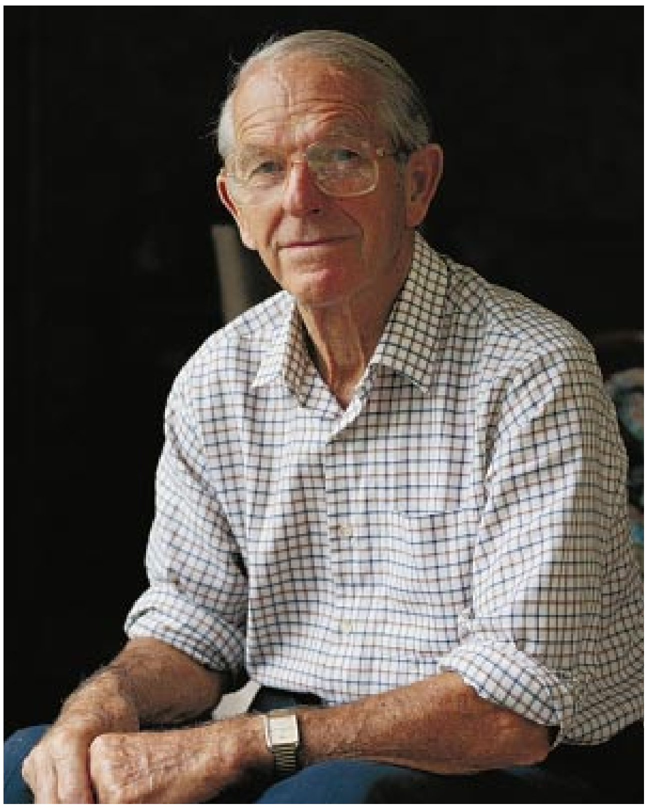
## Revealing the Hidden Sequence

One might expect a two-time Nobel Prize winner to spend his days raising funds for a world-class laboratory, giving lectures to adoring colleagues and collecting royalties on his best-selling novel. Yet Frederick Sanger seeks neither fame nor fortune. Instead the man who built the foundations of modern biochemistry lives quietly in Swaffham Bulbeck, England, tending a garden of daffodils, plum trees and herbs. “I think Sanger hasn’t been recognized as much as some, partly because he is an undemonstrative person,” says Alan R. Coulson, who collaborated with Sanger for 16 years at the MRC Laboratory of Molecular Biology in Cambridge.

Forty years ago, Sanger was the first to reveal the complete structure of a protein. Then during the 1970s he developed one of the first techniques for reading the genetic code. “He deserved two Nobel Prizes,” says biochemist G. Nigel Godson of New York University Medical Center. “He single-handedly engineered two revolutions in biology.” In addition, Coulson boasts, Sanger deserves much of the credit for laying the groundwork for the Human Genome Project, the multinational effort to determine the entire sequence of nucleotides in human DNA.

Sanger, now 75, chose to meet me at one of his favorite pubs, the Red Lion, near Cambridge. He has mastered the art of understatement in both appearance and action—a no-frills kind of guy. Wearing a red sweater over a plaid shirt, he orders a chicken sandwich and a half-pint of lager. He speaks in a shy voice that is barely audible above the noise of the restaurant. Sanger is uncomfortable talking to journalists, but in the relaxed atmosphere of the pub he reveals some of what made him one of the great scientists of this century.

Sanger was born into a wealthy home. His grandfather made a fortune in the cotton trade and passed it on to his mother, Cicely. As a boy, Frederick was fascinated by nature, collecting everything from rocks to insects. His early ambition was to become a doctor, like his father, Frederick senior. Yet although the younger Sanger enjoyed learning about the science of medicine, he was not interested in the art of diagnosing and treating disease.



DAVID LEVENSON *Black Star*

***FREDERICK SANGER transformed biochemistry by developing methods for determining the structure of proteins and DNA.***

By all accounts, Sanger was not a brilliant student. In 1936 he was accepted to the University of Cambridge, but he struggled with the basic sciences. Sanger received passing grades in chemistry, but he “bombed” in physics. “I never won scholarships,” he notes. “I am not sure I would have been able to attend Cambridge if my parents had not been fairly rich.”

Sanger found his calling early in his college career. An enthusiastic professor, Ernest Baldwin, enticed him to study the new science of biochemistry. Sanger enjoyed the subject so thoroughly that he took advanced courses during a fourth, extra year at Cambridge. Two weeks after the final exams, he says, he was “very surprised to learn that he had been awarded a first-class degree.”

In 1940, unlike most of his peers, the 22-year-old Sanger did not go off to fight in World War II. “I was brought up as a Quaker, and I felt pretty strongly that people should not go around killing others for any reason,” he explains. He successfully defended his position before a military tribunal and spent the war years pursuing a doctorate. Graduate students were obviously in short supply during the war, and the biochemistry department at Cambridge was all too happy to accept a promising student.

After earning his Ph.D. in 1943, Sanger joined the laboratory of Cambridge professor A. C. Chibnall, who was a pioneer in the field of protein chemistry. Chibnall asked Sanger to study insulin, the pancreatic hormone that governs the metabolism of sugar. The suggestion led to a 10-year-long project that established Sanger as the leader of his field. Insulin, like other proteins, is made up of different amino acids. During the 1940s researchers were very aware that the chemistry of proteins depends on the order in which the amino acids are arranged. The problem: no one had found a technique for deducing the sequence.

Chibnall and Sanger chose to work on insulin for several reasons. It was available in a pure form, and it was a small molecule, at least as proteins go. More important, perhaps, investigators realized that if they could work out the structure of insulin, they would understand how it controlled sugar metabolism—insights that had many implications for medicine.

To determine the sequence of insulin, Sanger began with a simple strategy that chemists often use to analyze large compounds: trying to break the molecule into fragments and then figure out how the pieces fit together. Easier said than done. To snip insulin into pieces of a meaningful size and then sort them, Sanger tried every trick in the chemist's handbook and then came up with some new schemes.

One of Sanger's important innovations was a method for labeling an end of a protein fragment. These tags made it easier to deduce what pieces belonged together. For example, if three amino acids—call them A, B and C—are linked together in some order in a chain, the sequence could be determined in the following way. After labeling the chains, they could be broken down into individual amino acids, indicating, say, that B was at one end of the chain. Then a new sample of the same chains could be cut into pieces consisting of two amino acids. The final step was isolating all those pieces that had a B at one end. If all the B's were linked to A's, then the sequence would have to be BAC.

While pursuing the insulin sequence, Sanger tried hundreds of different techniques. "Most experiments go wrong," Sanger sighs. "I didn't spend too much time trying to figure out what went wrong. I just started thinking about the next experiment. It prevented me from getting depressed."

There was one experiment in particular that Sanger would like to forget. In 1947 he spent a year in Uppsala, Sweden, working in the laboratory of the eminent biochemist Arne W. K. Tiselius. A technique Sanger developed seemed to suggest that insulin was not a single chain but rather four cross-linked chains. When he presented the results to his mentor, Tiselius suggested that they rush to publish a paper together. "I was rather shocked as he had not really contributed anything."

But as a junior member of the laboratory, Sanger gave in. "The paper is the only one of which I am ashamed," he comments. Sanger later discovered that insulin is actually made of two cross-linked chains, one 30 amino acids long and the other 21 long. The larger chain was by far the most complex structure that a protein chemist of that era had ever struggled with. It was not until 1952 that Sanger and his co-workers Hans Tuppy and E.O.P. Thompson figured out the complete sequence of both chains.

Sanger then had to decide how the chains linked together to make an insulin molecule. The problem turned out to be extremely complicated. An initial analysis seemed to show that chains were linked at nearly every point. Then Sanger realized that his techniques for cutting the insulin molecule were introducing new bridges between the strands. By 1955 he found a way to prevent the introduction of cross-links and succeeded in determining the complete structure of insulin.

Four years later, Sanger won the Nobel Prize in Chemistry for the insulin work. He was immediately besieged by professors inviting him to teach and administrators asking him for advice. But Sanger wanted no part of it. "I have actively tried to avoid both teaching and administrative work," he says. "This was partly because I thought I would be no good at them but also out of selfishness."

*Sanger's talent for solving the right problem at the right time earned him two Nobel prizes in chemistry.*

Ironically, in the year Sanger became a Nobel laureate, his research began to founder. He had taken the insulin study as far as it could go and was looking for new questions to grapple with. "I think these periods occur in most people's research careers and can be depressing and sometimes lead to disillusion," he observes. "I have found that the best antidote is to keep looking ahead."

In 1961 Sanger joined the Laboratory of Molecular Biology at the Medical Research Council in Cambridge and hired Coulson as his research assistant. "He's not the kind of guy you pal around with right away," Coulson says.

Soon after his arrival at the laboratory, Sanger decided his talents might be useful for analyzing DNA, the molecule that stores the genetic code. In the 1950s James Watson and Francis Crick had figured out that DNA was a long, double helix made of four different nucleotides. The arrangement of nucleotides determines what proteins an organism can make, that is, what genes it will express. Yet at the time, scientists could determine the sequence of nucleotides for only a very small section of a DNA strand.

Sanger set an ambitious goal for himself: sequencing the many thousands of nucleotides in the DNA of a virus. Yet he and other biochemists soon found DNA more difficult to analyze than proteins for two reasons. First, they had less experience handling DNA than they did proteins. Second, DNA is made from only four basic building blocks, whereas proteins are constructed from some 20 different amino acids. Just as a puzzle with many similar pieces is more difficult to solve than one with many different pieces is, the sequence of DNA was more difficult to decode than the sequence of insulin was.

For more than 10 years, Sanger investigated techniques for sequencing DNA, competing with several laboratories around the world. Then in 1975, Sanger and Walter Gilbert of Harvard University, working independently of one another, developed methods for rapidly sequencing DNA. Their techniques could determine the arrangement of nucleotides in segments of DNA 200 or more units long in a few days. With earlier methods, the job would have taken years. The new technique enabled Sanger and his collaborators to determine the sequence of 5,375 nucleotides in a virus called $\phi$X174.

In 1980 Sanger won a second Nobel Prize, which he shared with Gilbert and Paul Berg of Stanford University. Berg had found a way to insert pieces of DNA from one organism into the DNA of another. His work launched the technology of recombinant DNA.

During his career, Sanger published about one major scientific paper every eight years, but his colleagues say each one is a classic in experimental biochemistry. "I don't publish papers unless I have something to write about and something I am sure about," he explains.

These days, although Sanger still gives advice to his colleagues at the Laboratory of Molecular Biology, he has retired from his research. "The aging process was not improving my performance in the laboratory," he complains. "I would have felt guilty about occupying a space that could have been available to a younger person."

Reflecting on his research, Sanger finds it hard to recall any moments of great inspiration. He does not subscribe to the "popular idea that scientific progress depends on sudden breakthroughs." Instead he thinks about many events that "were more often associated with small and gradual advances."

Sanger finishes his lunch and invites me to his home. We drive through the countryside to an unpretentious house surrounded by an elaborate garden. It seems Sanger is as devoted to gardening as he once was to research. He and his wife of 52 years, Joan, grow a dozen varieties of flowers, cultivate several fruit trees and raise chickens. When they are not working in the garden, the Sangers entertain their three children, grandchildren and friends. "I have always had a happy and peaceful life," he remarks. The key, he asserts, "is working on the right thing at the right time." *—Russell Ruthen*

SCIENTIFIC AMERICAN

# Clearing the Air in Los Angeles

*Although Los Angeles has the most polluted skies in the nation, it is one of the few cities where air quality has improved in recent decades*

by James M. Lents and William J. Kelly

On some hot, sunny days, the 14 million residents of the Los Angeles area inhale a thick, brownish-gray haze, and no one can ignore its effect. The smog obscures the San Bernardino Mountains and the warm California sun; it irritates the eyes and nose; it restricts the activities of athletes and people who have breathing disorders; it injures the lungs of both young and old.

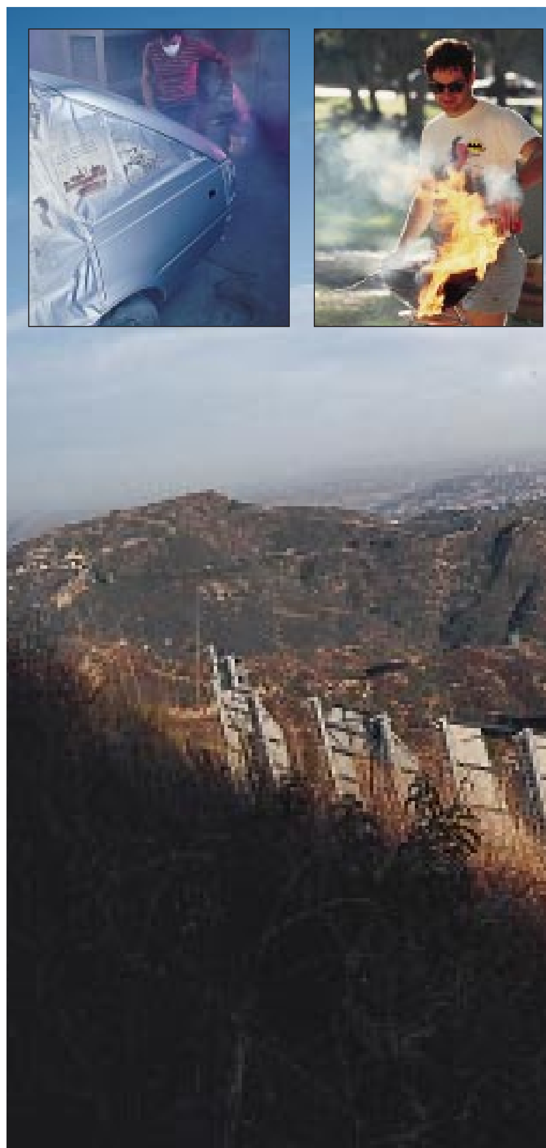
Southern California's air quality is the worst in the U.S. Air pollution in the region reaches unhealthful levels on half the days each year, and it violates four of the six federal standards for healthful air—those for ozone, fine particulates, carbon monoxide and nitrogen dioxide. In 1991 the South Coast Air Basin exceeded one or more federal health standards on 184 days.

Yet these statistics hide a remarkable accomplishment of the citizens of southern California. Los Angeles is one of the few places in the nation where air quality has improved dramatically since the 1970s. From 1955 to 1992 the peak level of ozone—one of the best indicators of air pollution—declined from 680 parts per billion to 300 parts per billion. The California Air Resources Board recently documented that population exposure to unhealthful ozone levels has been cut in half in just the past decade. Furthermore, the smog levels measured during each of the past three years have been the lowest on record.

All these improvements were achieved at a time when human activity in the Los Angeles area was increasing at a rapid rate. Since the 1950s the population has almost tripled, from 4.8 million to 14 million; the number of motor vehicles on the road has more than quadrupled, from 2.3 million to 10.6 million; and the city has grown into one of the most prosperous regions of the world.

Although the residents of southern

JAMES M. LENTS and WILLIAM J. KELLY work together at the South Coast Air Quality Management District (AQMD), the regional air-pollution control agency for the Greater Los Angeles area. Before becoming executive officer of AQMD in 1986, Lents headed the air-pollution control program for the state of Colorado. In 1970 he received a Ph.D. in physics from the Space Institute at the University of Tennessee. Some 11 years ago Kelly earned an M.A. in journalism from Columbia University, and since then he has written extensively on the environment.



**AIR POLLUTION SOURCES have increased in size and number in Los Angeles, yet technical innovation and social policy have led to an improvement in air quality during the past two decades. Some typical sources of air pollution include (*from left to right*) industrial coatings, barbecues, trash incinerators, paints, dry cleaners, commercial ovens and motor vehicles.**

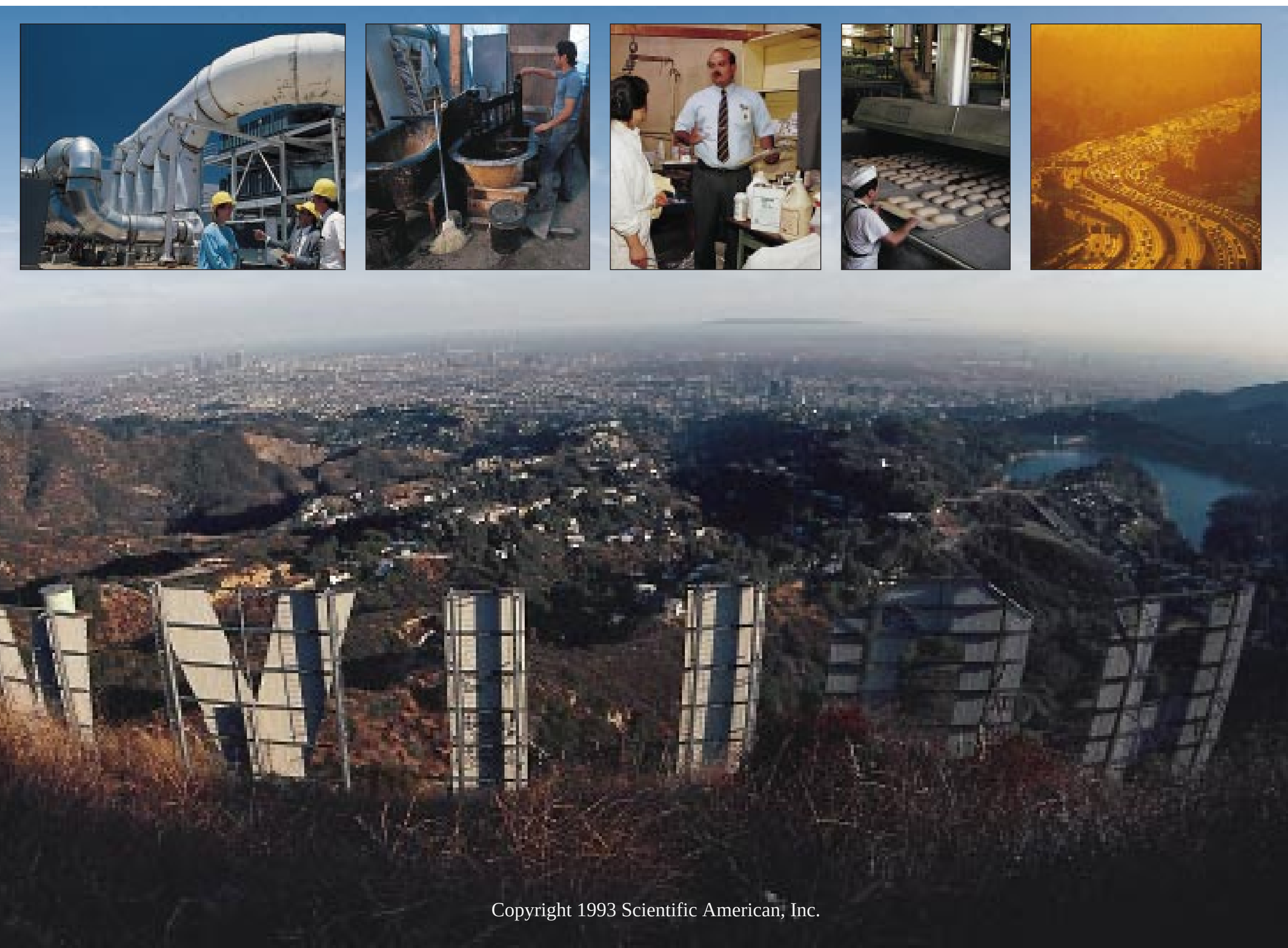
California still face and continue to tackle many air pollution problems, they have an advantage in that they have spent 50 years studying the local atmosphere and experimenting with various policies. We hope the cities of all nations will learn from the experiences of Los Angeles [see "The Changing Atmosphere," by Thomas E. Graedel and Paul J. Crutzen; SCIENTIFIC AMERICAN, September 1989].

The movement to clean up the air in southern California began during the 1940s, a period of rapid industrialization. At the time, the region was plagued by sudden "gas attacks" that irritated the eyes, diminished visibility and produced an unpleasant odor. Then, as now, the smog was so obvious and odious to the public that elected leaders were compelled to take meaningful action. Yet their efforts provoked strong conflict. Some citizens and industries fiercely resisted suggestions to clean up sources of pollution. But the *Los Angeles Times* published dozens of editorials demanding that the smog problem be solved. The paper also put its money behind its editorial mouth. In 1947 it retained Raymond R. Tucker, the former smoke regulation commissioner of St. Louis, to study air pollution in the area. Tucker identified and investigated several major sources of air pollution, including heavy industries, foundries, motor vehicles, backyard incinerators and smudge pots for protecting crops from frost.

In the same year, the oil industry paid the Stanford Research Institute (SRI) to give another perspective on the causes and control of pollution. The organization discovered that the hazy days were caused in part by a natural weather phenomenon known as an inversion layer. The warmest part of the atmosphere is, more often than not, that nearest to the ground, but under certain conditions a layer of cool air can slip underneath a stratum of warm air. Such inversions often form off the coast of Los Angeles as the Pacific Ocean cools the atmosphere just above it. After ocean breezes blow the air mass inland, the inversion layer traps air pollutants in the cool air near the ground where people live and breathe. The mountains that surround the region compound the problem; they prevent the pollutants from dispersing.

SRI pointed out that natural materials such as dust, pollen, fibers and salt were important components of the haze. But the institute also recognized that industries and motor vehicles contributed to the problem by adding carbon particles, metallic dust, oil droplets and water vapor.

In the 1950s SRI and Arie J. Haagen-Smit and his colleagues at the California Institute of Technology began to examine the chemistry of the atmosphere above Los Angeles. Their work and the research of others have revealed the complexity of atmospheric chemistry. Automobiles, factories and other sources release such raw pollutants as hydrocarbons, water vapor, carbon monoxide and heavy metals. When these chemicals are exposed to intense sunshine, they react to yield a vast number of secondary pollutants—for instance, ozone, nitrogen dioxide, various organ-



ic compounds and acidic particles of nitrate and sulfate. This concoction then interacts with plants and animals, causing a variety of different effects. Many of these phenomena are still not understood, but 50 years ago even less information was available.

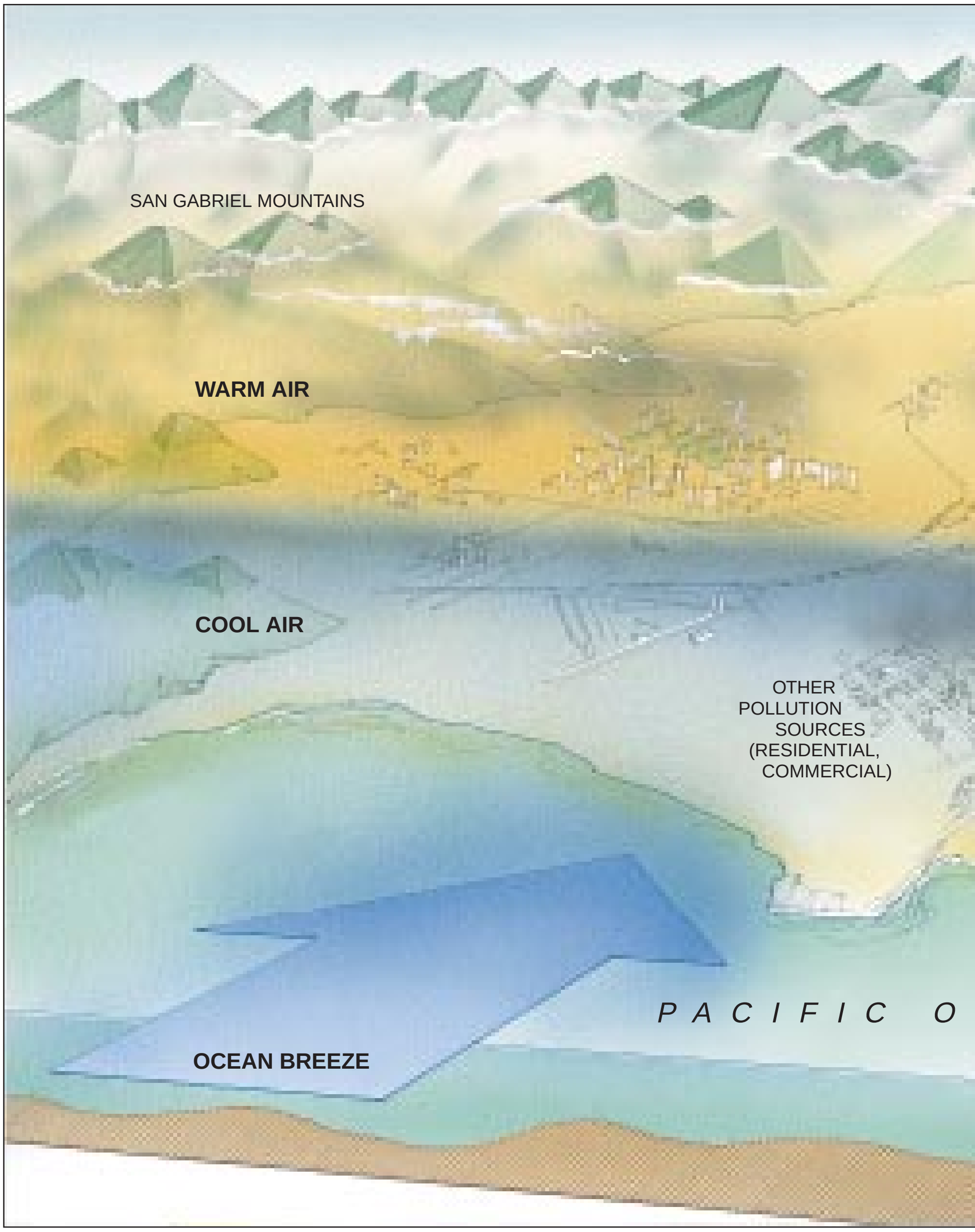
In 1953, with the public fearing that the Los Angeles haze might become as bad as London's "killer" fog, Governor Goodwin J. Knight appointed an air-pollution review committee. Chaired by Arnold O. Beckman of Beckman Instruments, the committee proposed five key ideas for reducing pollution over the short term. First, they asked that the emission of hydrocarbons be reduced by improving procedures for transferring petroleum products. Second, they set standards for automobile exhausts. Third, they encouraged the use of trucks and buses that burned liquefied petroleum gas instead of diesel fuels. Fourth, they considered whether industries that polluted the area heavily should be asked to slow their growth. Fifth, they advocated that the open burning of trash be banned.

The committee also hoped that over the long term Los Angeles would develop a sustained automotive pollution-control program, construct a rapid transit system and start a cooperative program to regulate industrial sources of pollution. Ironically, the report was issued while the region's public train system was being dismantled. Today Los Angeles is trying to get back on track, so to speak, by developing an extensive regional commuter rail network. The Beckman committee's recommendations eventually grew into a coherent air-quality management plan for the region, but it emerged slowly and was reshaped many times.

Soon after the Beckman report, the neighboring counties of Orange, Riverside and San Bernardino began their own pollution-control programs. During the past 40 years, these regions have experienced explosive growth in population and in vehicular traffic. The control programs in these regions were, for the most part, as energetic and innovative as those in Los Angeles. But residents of these counties soon realized that they needed to coordinate their efforts; smog does not respect political boundaries.

In 1975 the regional governments tried voluntarily to consolidate their pollution-control programs. Two years later little progress had been made. So the California legislature forged an uneasy alliance between the local programs by creating the South Coast Air Quality Management District (AQMD). The AQMD was given jurisdiction over the counties of Los Angeles, Orange and Riverside and part of San Bernardino—an area of 13,350 square miles.

Initially, the AQMD was responsible for stationary sources of air pollution, and the California Air Resources Board was assigned to regulate mobile sources such as cars, trucks and buses. In its early years the AQMD adopted a viewpoint held by many business leaders. It argued that many industries were as clean as they could get, the state was not doing enough to clean up cars and the region might never achieve clean air standards. As required by law, the AQMD adopted air-quality management plans in 1979 and 1982, but these were regarded as mostly paper exercises. At the time, the federal Clean Air Act required all American cities to achieve federal standards by 1987, although nearly everyone realized that the task would be impossible in Los Angeles. As that deadline drew near, however, environmentalists and even some business groups attacked the AQMD for its complacency and alleged lax enforcement. In 1987 the legislature restructured the AQMD governing board and granted broad powers to the new



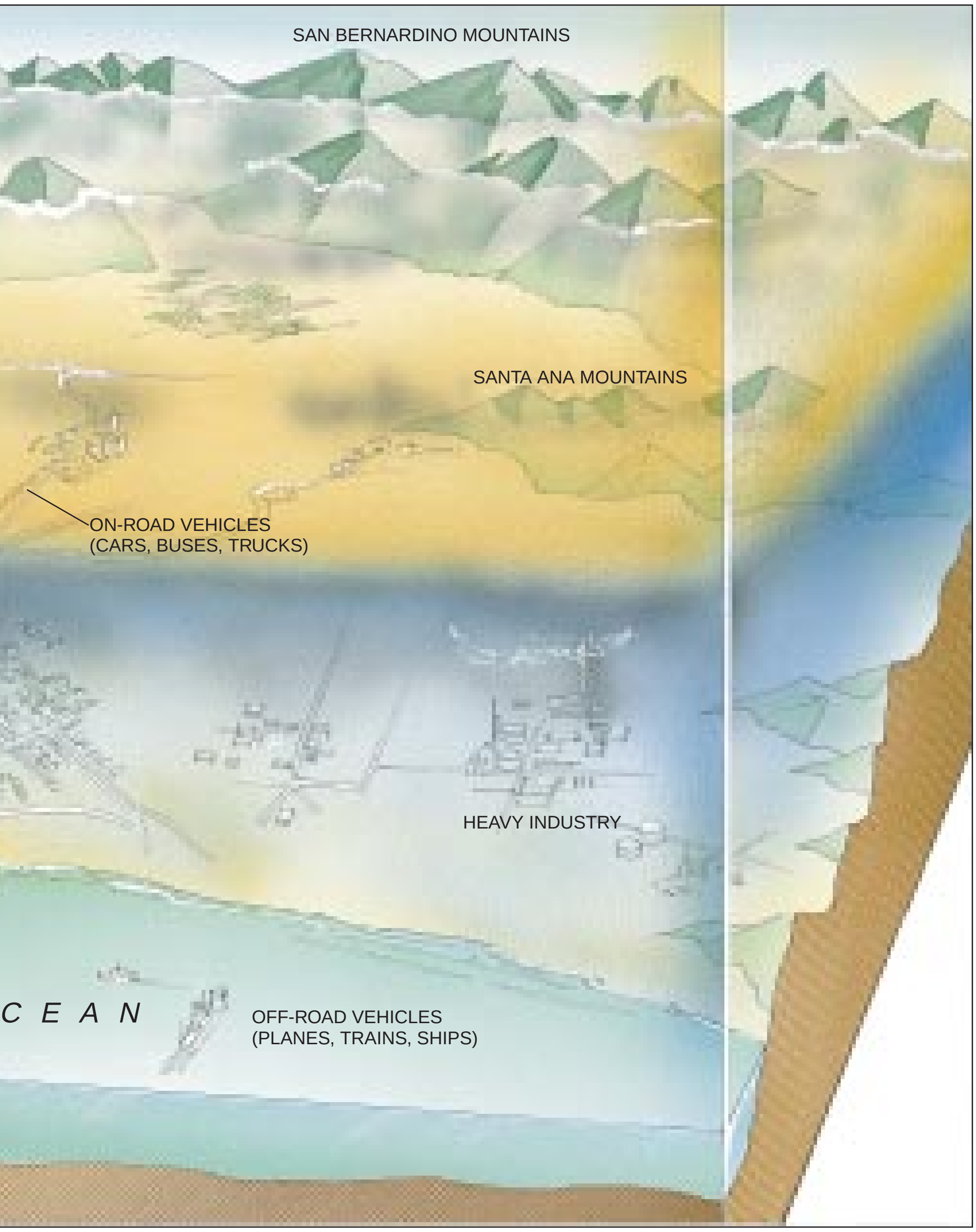

**WEATHER PATTERN known as an inversion layer traps air pollution above Los Angeles and neighboring counties. Air cools over the ocean and then is blown inland, creating a cold layer near the ground and a warmer layer above. The warm layer prevents most of the smog from escaping upward. The surrounding mountains keep the polluted air from moving farther inland.**

board. The AQMD now has responsibility for achieving local, state and federal standards.

Despite all the political upheaval, the state and local governments had managed to curb many sources of air pollution in the South Coast Air Basin before the AQMD was formed in 1977. They gradually developed regulations that concentrated on reducing the major sources of air pollution: particles from trash incineration, emissions from industry and pollutants from motor vehicles.

In 1958, at the recommendation of the Beckman committee, backyard trash incinerators were banned despite opposition from the public and some waste-disposal managers. More than 300,000 families owned such incinerators, and many were unwilling to give up the convenience and cost savings. But slowly attitudes changed; today most residents of Los Angeles would consider it shameful to burn their trash and subject their neighbors to the smoke and smell.

Waste-disposal managers objected to the ban for a different reason. These analysts realized that the economical alternative to burning trash was burying it, and therefore they correctly predicted that although replacing incinerators with landfills would reduce air pollution, the additional landfills would create other types of environmental problems. For example, as rain has seeped through the landfills and carried away soluble materials, it has contaminated the local groundwater. Southern Californians are now working to clean up the groundwater, but the long-term plans are, first, to decrease the volume of trash through recycling and, second, to reduce further groundwater contamination by improving the management of landfills.

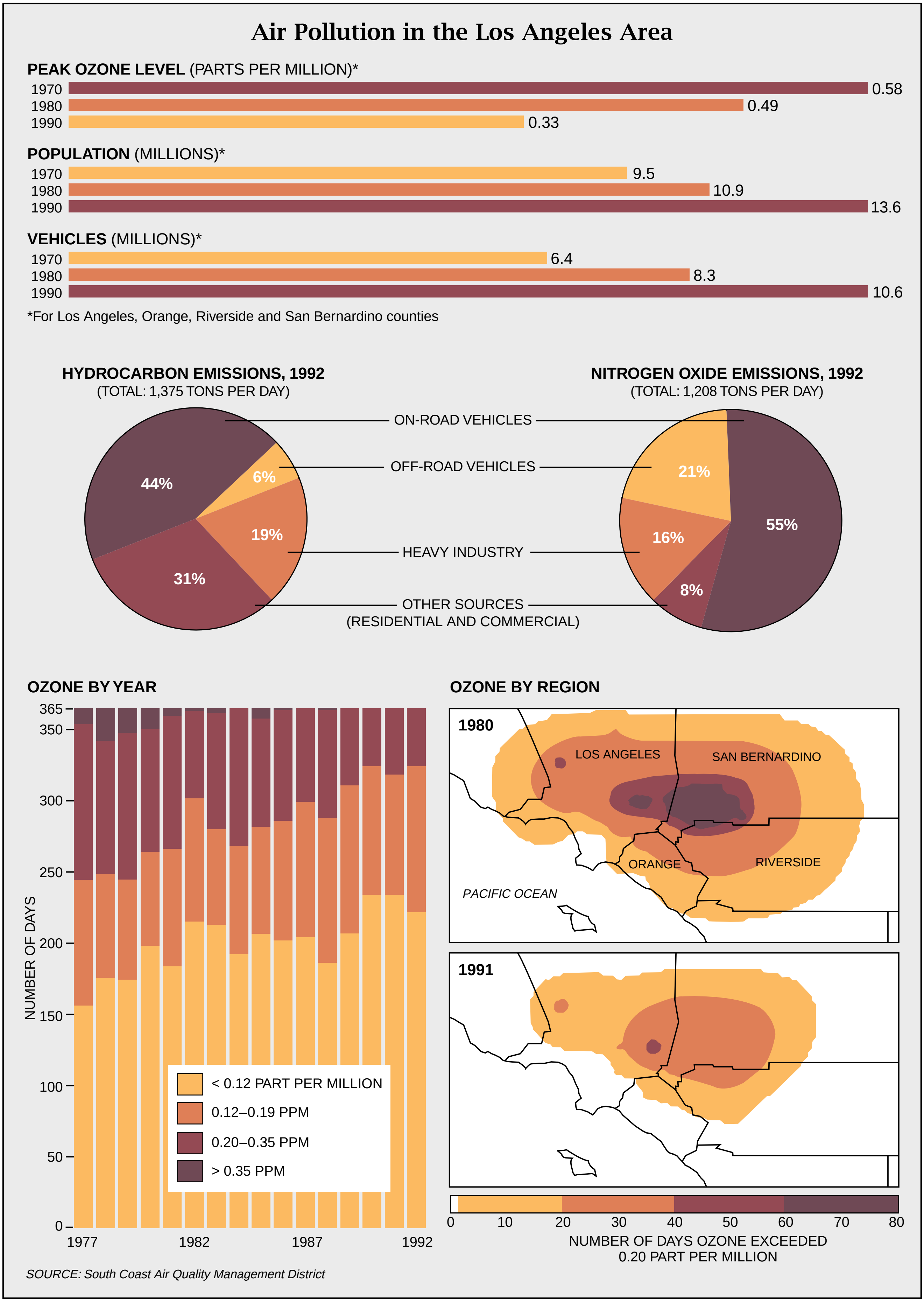



The ban of backyard incinerators generated relatively little public resistance when compared with attempts to clean up industrial sources of pollution. Historically, the control of industrial emissions has been challenging because of the need to balance the region's environmental interests with its economic needs. The Beckman report of 1953 quickly led to requirements that industries use vapor-recovery equipment when they transferred petroleum products, but service stations were not obliged to install such equipment on gasoline pumps until 1978. This equipment, while originally somewhat difficult for the motorist to handle, has been streamlined, so it is effective and easy to use today. Yet the measure is still resisted by most communities outside of California because of its expense and an undeserved reputation for being cumbersome.

Regulations were adopted during the 1960s to eliminate industrial solvents that play a major role in promoting the formation of ozone. The rules affected a wide variety of businesses, from construction to auto manufacturing to dry cleaning. To meet the requirements, most industries chose to use nonreactive solvents instead of installing control equipment because it was the less expensive solution. Sadly, the nonreactive solvents were later shown to destroy ozone in the stratosphere. (At high altitudes, ozone serves the critical function of shielding the earth from harmful solar radiation.) The solution to this problem is to find truly benign solvents; for example, Hughes Aircraft Company has recently developed a soldering flux made from citrus juice.

Although Californians succeeded in reducing the quantity of pollution generated by certain industrial sources, their efforts have been partially offset in recent decades. Many sources have been introduced as the region's economy has grown. To compensate for the growth factor, officials instituted, in 1976, "new source" regulations. These required expanding industries to use the cleanest technology available. The rules also specified that if a company planned to start a project that increased emissions, it was required to earn a certain number of credits by reducing emissions from another project.

The new source regulations have stim-


Air Pollution in the Los Angeles Area
PEAK OZONE LEVEL (PARTS PER MILLION)*
1970 0.58
1980 0.49
1990 0.33
POPULATION (MILLIONS)*
1970 9.5
1980 10.9
1990 13.6
VEHICLES (MILLIONS)*
1970 6.4
1980 8.3
1990 10.6
*For Los Angeles, Orange, Riverside and San Bernardino counties
HYDROCARBON EMISSIONS, 1992
(TOTAL: 1,375 TONS PER DAY)
NITROGEN OXIDE EMISSIONS, 1992
(TOTAL: 1,208 TONS PER DAY)
ON-ROAD VEHICLES
OFF-ROAD VEHICLES
HEAVY INDUSTRY
OTHER SOURCES
(RESIDENTIAL AND COMMERCIAL)
44%
6%
19%
31%
21%
55%
16%
8%
OZONE BY YEAR
NUMBER OF DAYS
365 350 300 250 200 150 100 50 0
1977 1982 1987 1992
< 0.12 PART PER MILLION
0.12–0.19 PPM
0.20–0.35 PPM
> 0.35 PPM
OZONE BY REGION
1980
LOS ANGELES
SAN BERNARDINO
ORANGE
RIVERSIDE
PACIFIC OCEAN
1991
0 10 20 30 40 50 60 70 80
NUMBER OF DAYS OZONE EXCEEDED
0.20 PART PER MILLION
SOURCE: South Coast Air Quality Management District

ulated considerable innovation. When researchers compared the best control technologies of 1976 with those available in 1990, they found that the amount of pollution generated during the manufacture of a product had dropped, on average, by 80 percent. Unfortunately, many businesses found a way to avoid the regulations. The rules exempted new projects that would emit daily less than 75 pounds of hydrocarbons or 100 pounds of nitrogen oxides. To exploit this loophole, many industries expanded by adding several small projects instead of a few large ones so that all the undertakings would be exempt. Furthermore, most small businesses were exempt, even though they have accounted for the fastest-growing segment of stationary source emissions during the past few decades.

Growth of these small sources, combined with increased population and motor vehicle use, tended to offset cleanup efforts through the 1980s. Consequently, in 1990 the new source regulations were revised to include projects that introduced small quantities of air pollution. Even then, some exemptions were allowed, and they continue to hurt the effort.

Los Angeles's difficulties with reducing industrial pollutants were similar in one respect to its attempts at diminishing automotive emissions. Although officials identified the problem and took action, they were slow to adjust to the scale of the matter. From the start, for example, the Beckman committee recommended controls on automotive emissions, but it did not anticipate how quickly auto travel would grow throughout the region. Nevertheless, Los Angeles has succeeded in three important areas: promoting the use of clean fuels, purifying engine exhaust and encouraging carpooling and the use of public transportation.

A significant first step in regulating automotive pollution was the reduction of "blowby" gases. Most car engines that were designed 30 years ago allowed the gaseous by-products of combustion to escape past the piston, through the crankcase and into the atmosphere. In 1960 such blowby gases accounted for one quarter of the total hydrocarbon emissions in the region. The gases could be eliminated by installing a $5 device that routed the crankcase emissions to the engine intake manifold; in this way, the hydrocarbon fumes were burned instead of being released. In 1963 the state mandated that the devices be added to all motor vehicles, new and old. By the middle of the decade, a special police force was patrolling the roadways in southern California, issuing citations to vehicles that smoked excessively.

The crankcase device proved effective on new vehicles, but objections were raised to equipping old cars with them. Many auto mechanics were not trained to install the devices, and rumors spread that even properly installed devices caused engine damage. That misinformation was the primary reason for the state legislature's decision in 1965 to stop requiring the installation of crankcase devices in cars made before 1963. It was not the first time, nor certainly the last, that inaccurate information has become the basis of political compromise.

In addition to installation of crankcase devices, Los Angeles tried to reduce automotive pollution by encouraging the production and use of clean gasolines. In 1960 officials required that all gasoline sold in the area have a low content of olefins, which are some of the most reactive compounds in petroleum products. Unfortunately, as the years passed by, the clean gasoline program lost its momentum. Then, in the early 1980s, the state government of Colorado revitalized the concept, hoping to reduce air pollution problems in Denver. Colorado required the blending of a methanol derivative into gasoline. In California, meanwhile, the AQMD and the state were pushing for methanol and other alternative fuels through regulatory programs. Faced with losing large market shares to methanol and other clean fuels, the Atlantic Richfield Company in Los Angeles came up with a constructive solution by developing a truly clean gasoline. These events stimulated California and the federal government to go beyond regulations and introduce new legislation for clean fuels.

Another development in reducing automotive emissions was the removal of pollutants from the exhausts of car engines through the use of catalytic converters. California required that beginning with the 1975 model year, all new cars have the converters. The converters and similar devices substantially reduced emissions, but for several reasons the program was not as successful as it might have been. First, researchers discovered that the performance of a converter decreases as the car accumulates mileage. Second, many motorists tampered with or even removed the emissions-control equipment. Third, many consumers damaged their converters by fueling their cars with cheap leaded gasolines instead of the more expensive unleaded varieties. Indeed, in 1983 the National Enforcement Investigations Center found that in almost one fifth of all the vehicles on the road, the emissions-control equipment was either removed by the owner or damaged by the use of leaded fuels. The latter problem has abated in recent years as the price of unleaded fuels has moved closer to that of leaded gasoline.

To enforce legislation on automotive emissions, Los Angeles initiated an inspection program in 1976. The regulations specified that whenever a vehicle was sold, the new owner was required to bring the car to a special station where the emissions could be measured with the engine of the car running at normal cruising speeds. Los Angeles had many difficulties managing the program, and the state government was reluctant to help. Consequently, the city received innumerable complaints from car owners who were forced to wait in long lines to get their vehicles tested.

The California legislature was eventually persuaded to develop a new inspection program when the Environmental Protection Agency imposed sanctions that reduced highway funding. Under the new "smog check" program, each car in California is tested once every two years at a repair shop. Furthermore, the state gave local governments the option to conduct tests either while the engine was idling or while it was running at normal speeds—the latter option being more expensive but more effective. The program achieved pollution reductions, but the results were lower than expected. Today the federal government is again pressuring California to create a program that lives up to its potential.

Officials had much trouble encouraging southern Californians to share rides, thereby reducing congestion and emissions. One way to promote ride sharing is to reserve lanes on major highways for the use of cars transporting two or more people. When state transportation administrators proposed restricting one of four lanes of the Santa Monica Freeway, however, motorists opposed the idea so strenuously that the officials quickly withdrew it. Later they found that the public was less antagonistic if new lanes were added to existing highways for the use of high-occupancy vehicles. California now has an extensive network of such lanes.

In 1987 the AQMD required firms with 100 or more employees to offer incentives to institute carpooling. Although the regulation generated complaints, it has been successful. The rule applies to 5,200 work sites, harboring 1.2 million employees. A study of that group from 1987 to 1992 revealed that the number of employees per vehicle rose from 1.13 to 1.24. The program has eliminated 90,000 trips a day and has achieved about half the desired reductions in emissions. In fact, southern California is one of the few places in the U.S. where ride sharing has increased over the past five years.

Today the motor vehicles and industries operating in California are among the cleanest in the world. A new car sold in California emits just one tenth of the pollution that a new car did in 1970. Such industries as electric utilities rely almost exclusively on clean-burning natural gas. Manufacturing plants and construction companies use advanced paints, solvents and adhesives that have been formulated to minimize pollution. For these reasons and others, southern California has made tremendous progress in reducing air pollution.

But despite all the successes, pollution in the South Coast Air Basin is still overwhelming. Average daily emissions total 1,375 tons of hydrocarbons, 1,208 tons of nitrogen oxides, 4,987 tons of carbon monoxide, 134 tons of sulfur oxides and 1,075 tons of particulates. Transportation, including cars, airplanes, trains and ships, adds 47 percent of the hydrocarbons, 70 percent of the nitrogen oxides, 90 percent of the carbon monoxide, 60 percent of the sulfur oxides and 89 percent of the particulates. Industry contributes 26 percent of the hydrocarbons, 18 percent of the nitrogen oxides and 30 percent of the sulfur oxides and smaller percentages of other pollutants. The remainder of the pollutants are emitted by households and service-oriented businesses, such as restaurants, dry cleaners, gas stations and operations in commercial buildings.

Los Angeles must reduce pollution even further to meet federal health standards. Computer projections indicate that southern Californians must cut hydrocarbons by 80 percent, nitrogen oxides by 70 percent, sulfur oxides by 62 percent and particulates by 20 percent. The consequences of complacency could be disastrous.

Air pollution in the South Coast Air Basin has already taken a staggering toll on residents. A growing body of evidence reveals that the smog is a serious health hazard.

In 1991 David Abbey, an epidemiologist at Loma Linda University, found a correlation between long-term exposure to air pollution and the development of chronic diseases. Abbey studied 6,340 Seventh-Day Adventists, 62 percent of whom lived in the basin (the remainder lived throughout California). In that group of 6,340, those who resided in areas that exceeded government standards for suspended particles on 42 days or more per year had a higher risk of respiratory disease, including a 33 percent greater risk of bronchitis and a 74 percent greater risk of asthma. In addition, women living in those areas had a 37 percent greater risk of developing some form of cancer.

**MESSENGERS for the Rapid Blueprint Company in Los Angeles were outfitted with gas masks in the fall of 1955 so that they would not suffer from the effects of the smog. Shortly thereafter, local officials began air-pollution control programs.**

Roger Detels of the University of California at Los Angeles studied respiratory disease among residents in three areas of southern California that differ in air quality. During a five-year period, he periodically questioned and examined volunteers, aged seven to 59 years, to determine the health of their respiratory systems. In 1987 Detels discovered that residents who lived in high-pollution areas had more symptoms of respiratory disease, such as bronchitis and asthma, than did people who resided in low-pollution regions.

If the citizens of the South Coast Air Basin succeeded in meeting federal standards for ozone and particulates, a 1989 study by the AQMD predicted, they would gain $9.4 billion in health benefits every year. If the residents could reduce just particulates to the federal standard, they would prevent 1,600 premature deaths annually among those who suffer from chronic respiratory disease. In addition, they would eliminate 15 million person-days on which people with respiratory disease are unable to go to work, school and other activities, and they would reduce the risk of dying prematurely from exposure to particulates. (The risk of death is about one in 10,000—about half the risk a Californian faces of dying in a car accident.)

If Los Angeles residents were able to achieve the federal standard for ozone, they would eliminate annually 18 million person-days of restricted activity, 65 million person-days of chest discomfort, 100 million person-days of headache, 120 million person-days of coughing, 180 million person-days of sore throats and 190 million person-days of eye irritation.

These terrible statistics have captured the attention of the federal and state governments. The federal Clean Air Act, which was amended in 1990, gives Los Angeles until 2010 to achieve federal health standards, but the law also requires the region to make incremental progress toward healthful air. The California Clean Air Act demands that the region reduce emissions by 5 percent a year until health standards have been met.

Between 1989 and 1991 the AQMD devised an air-quality management plan to respond to the strong mandates of the federal and state law. During the next 17 years, the plan seeks to reduce pollution from virtually all sources and foster the development of new and cleaner technologies. The AQMD will execute the plan with the cooperation of the

EPA, local governments and state Air Resources Board. Each organization has a well-defined role. The AQMD is charged with cleaning up stationary sources and encouraging carpooling. The EPA is to set standards for airplanes, trains and ships that travel through the region. Local governments will work to alleviate traffic through improvements in the transportation infrastructure and expanding mass transit. The Air Resources Board is responsible for motor vehicle standards. It has already taken aggressive action that should reduce automotive emissions by 85 percent by the next decade and should increase sales of electric cars.

**L.A. MECHANIC at Green Motorworks, a local manufacturer of electric vehicles, converts a gas-fueled car to battery power. Such initiatives are receiving strong support from California state agencies as a way to meet ambitious air-quality standards.**

The AQMD and its sibling agencies plan to attack the pollution problem in three stages. The first stage, dubbed Tier I, includes 135 measures that can be accomplished using existing technologies and are to be adopted by 1996. The measures limit and reduce pollutants from such sources as electric utilities, motor vehicles, small businesses and even backyard barbecues. The electric companies will be installing low-polluting burners and catalysts on power plants to reduce nitrogen oxide emissions. Pollution from backyard barbecues will diminish as households use reformulated charcoal lighter fluid and other products for lighting grills. Automotive emissions will decrease because of tailpipe standards as well as programs encouraging carpools and use of public transportation.

The second stage, or Tier II, will take the region into the 21st century. These measures rely on technologies that have just entered the commercial market. The list includes a new house paint, developed by the Glidden Company, that does not release hydrocarbons. Another initiative is developing automobile engines that run on methanol, natural gas or other alternative fuels. The third stage, Tier III, requires technologies that have not been fully developed but are likely to be available in the next decade or so. Researchers are fabricating paint coatings that dry under ultraviolet lamps, without emitting any significant quantity of pollutants. They are also working on fuel cells, such as those used in the space shuttle, which will power a kind of zero-emission electric vehicle.

To help develop these new technologies, the AQMD is working with government agencies and private corporations through its technology advancement office. To date, the office has contributed some $40 million in seed money to facilitate almost 250 projects. Among the technologies now entering commercialization are alternative-fuel vehicles, industrial emissions controls for nitrogen oxides, and cleaner paints, solvents and coatings.

After the AQMD released its ambitious plan to clean up the air, the organization quickly realized it could not simply play the role of tough pollution cop and expect companies, especially small ones, to follow the plan. The region's businesses, which were caught in a recession and faced cuts in defense spending, were reluctant to invest resources in pollution-control equipment.

To sustain progress, a different approach was required, especially to clean up the growing number of small sources of air pollution. In 1991 the AQMD outlined a series of reforms, which included providing free technical assistance and starting a guaranteed loan program so that small businesses could purchase pollution-control equipment.

In the long run the task of getting businesses to reduce pollution will hinge on technological innovation. One of the AQMD's most powerful incentives for developing these clean technologies is the Regional Clean Air Incentives Market Program. Under this program, a business is asked to meet emission standards not only for individual pieces of equipment but also for its entire operations and facilities. Those businesses that cut pollution below the standard will be granted emission-reduction credits. Those companies that generate more pollution than the standard must buy enough credits to make up the difference. The program gives firms flexibility to choose how to reduce emissions, and it also provides an incentive for innovation because firms can make a profit by cleaning up the air.

During the past 40 years, the citizens of Los Angeles have been very active in the fight for clean air, but from time to time they have neglected the issue because they feared that air-quality regulations were incompatible with the region's economic interests. Today, despite occasional doubts about environmental regulation and legitimate concerns about the post–cold war economy of Los Angeles, residents and business people seem to recognize the need to solve the serious air-pollution problems.

We hope that by 2010 when our children climb the San Bernardino Mountains, they will see not a sea of smog but rather the Pacific Ocean and Catalina Island. We would like them to live in a region where technological innovation sustains both the economy and the environment. But most important, we want them to reap the benefits of healthy, fresh air.

FURTHER READING

ENVIRONMENTAL ETHICS: DUTIES TO AND VALUES IN THE NATURAL WORLD. Holmes Rolston. Temple University Press, 1989.

ONE EARTH, ONE FUTURE: OUR CHANGING GLOBAL ENVIRONMENT. Cheryl S. Silver and Ruth DeFries. National Academy Press, 1990.

COSTING THE EARTH: THE CHALLENGE FOR GOVERNMENTS, THE OPPORTUNITIES FOR BUSINESS. Frances Cairncross. Harvard Business School Press, 1992.

EARTH IN THE BALANCE: ECOLOGY AND THE HUMAN SPIRIT. Al Gore. Houghton Mifflin, 1992.

# Large Igneous Provinces

*These vast fields of lava record powerful but geologically brief pulses of magmatic activity. Their formation may have triggered significant changes in the global environment*

by Millard F. Coffin and Olav Eldholm

Sometimes getting there is not half the fun of geophysical research. Bearing westward into a rising Southern Ocean gale, the Australian research vessel *Rig Seismic* was pounded so hard that its anchor broke loose and smashed into the forecastle, causing the waves breaking over the bow to flood parts of the ship. Progress toward the Kerguelen Plateau, an elevated region of seafloor just north of Antarctica, was delayed while the ship's crew secured the anchor and welded the ship to keep it watertight. Several days later the expedition, including one of us (Coffin), arrived at Kerguelen, deployed instruments for analyzing the structure of the ocean bottom and began collecting data. The aim of the voyage was to understand the origin and evolution of the huge underwater plateau.

In that same year, 1985, but in the other hemisphere, a floating earth science laboratory and drilling vessel, *JOIDES Resolution,* steamed north through stormy North Atlantic seas to the Vøring Plateau off the coast of Norway. That expedition, led by one of us (Eldholm) and Jörn Thiede of the Research Center for Maritime Earth Sciences in Kiel, Germany, had as its purpose the investigation of the geologic structures that form when continents tear apart and an ocean is born. Confounding the expectations of most of their colleagues, the researchers managed to obtain core samples of igneous rock nearly one kilometer deep, lying under 300 meters of soft sediment.

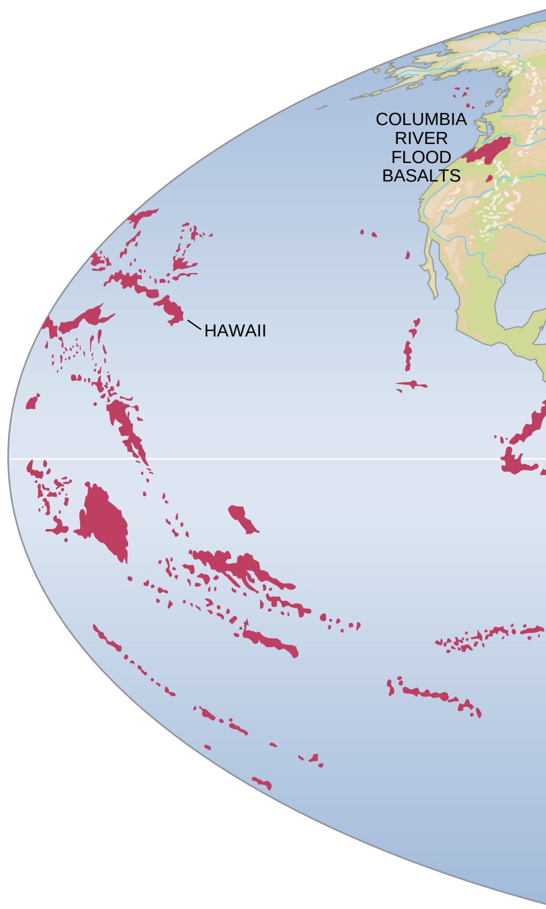
Although they lie thousands of kilometers apart in disparate geologic settings, we recognized fundamental similarities in the two plateaus. Our findings, merged with seismic and drilling data gathered by many other workers, have helped demonstrate that the Kerguelen Plateau and the volcanic continental margin off Norway belong to a class of huge magmatic features generically known as large igneous provinces. These features cover areas of up to millions of square kilometers, and yet they seem to have formed quite swiftly in geologic terms.

The spasms of eruptive activity associated with such rapid outpourings of lava may have substantially affected the chemistry and circulation of the oceans and atmosphere. Some of the resulting environmental shifts may have contributed to mass extinctions, including the one in which the dinosaurs vanished; in contrast, other changes may have promoted biological diversity and the origin of new species.

The recognition of large igneous provinces has forced geophysicists to rethink their notions regarding the structure of the earth's interior. The theory of plate tectonics tidily accounts for the slow, steady volcanic activity that occurs at mid-ocean ridges (where new oceanic crust is born) and near subduction zones (where old, dense sections of the ocean floor sink back into the earth's hot interior), but it cannot readily explain the abrupt outbursts necessary to create large igneous provinces. Although the rocks that make up these features generally resemble the composition of the lavas that emerge at mid-ocean ridges, igneous provinces differ in their trace element content and mix of atomic isotopes. And they sometimes break through the middle of the normally placid lithospheric plates.

In the 1960s the late J. Tuzo Wilson of the University of Toronto, followed by W. Jason Morgan of Princeton University, developed a hypothesis that helped

MILLARD F. COFFIN and OLAV ELDHOLM have collaborated for the past three years on a project to develop a fuller understanding of giant igneous formations. Coffin is a research scientist at the Institute of Geophysics at the University of Texas at Austin. He received his Ph.D. in geology in 1985 from Columbia University; he met Eldholm while completing his doctorate work. Eldholm is a professor of marine geophysics at the University of Oslo in Norway, a position he has held since 1981. He completed his doctoral work in geophysics at the University of Bergen in 1976. In 1991 he took a sabbatical at the University of Texas to coordinate his investigations with Coffin's; a year later Coffin took a fellowship in Norway. Both men have spent considerable time on board research vessels collecting data on submarine igneous provinces.




**LARGE IGNEOUS PROVINCES (*red regions*) appear in widely varied geologic settings around the globe. Dots indicate**

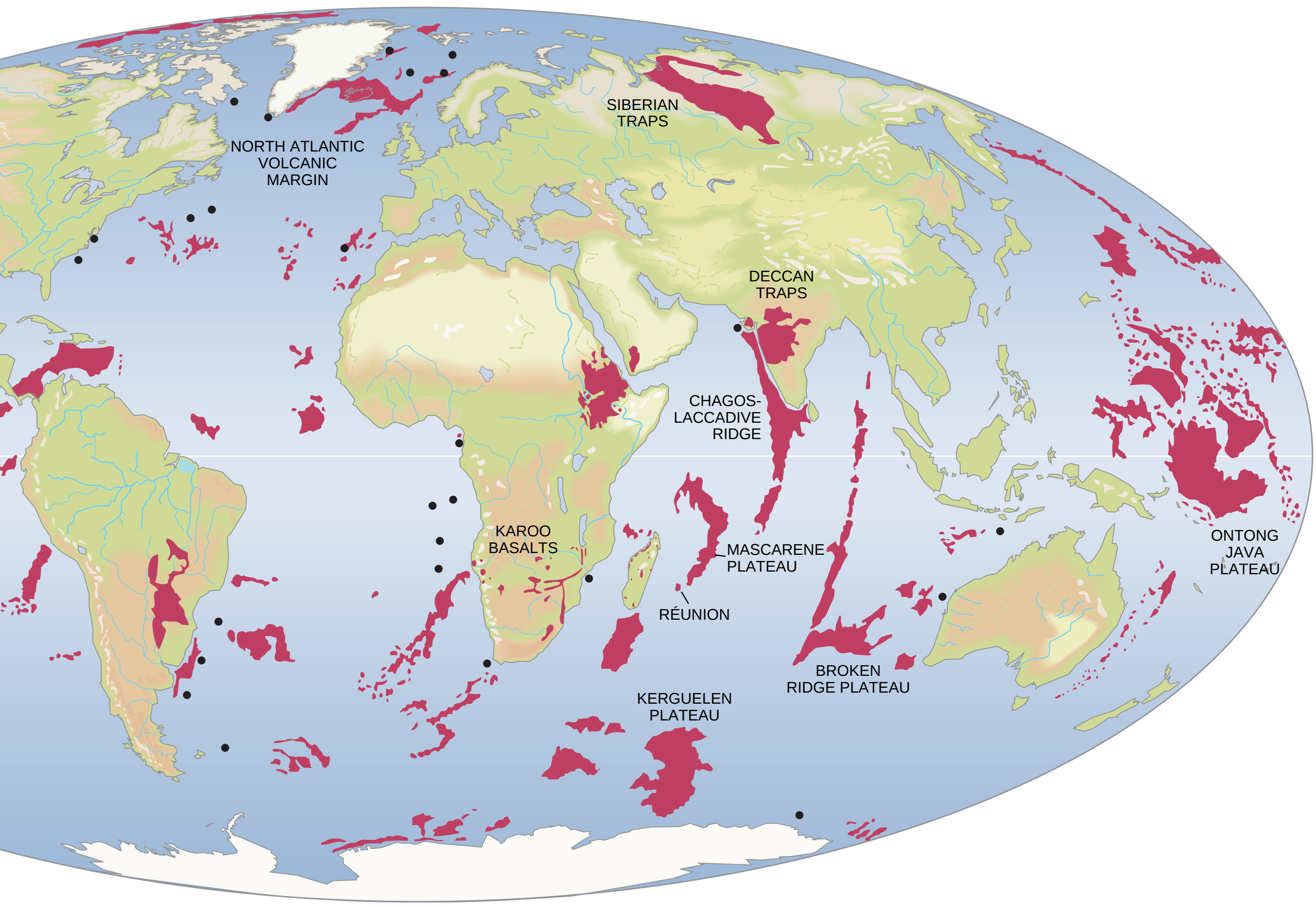
to elucidate the phenomenon. The researchers proposed that the earth's mantle, the vast zone lying below the crust but above the core, circulates in two different modes. The dominant one consists of large-scale convection that nudges plates across the surface and causes continents to drift. But about one tenth of the heat now escaping from the mantle does so in the form of deep-rooted, narrow plumes of warmer than average material that rise through the mantle. When it reaches the base of the lithosphere, a plume decompresses and partially melts, producing an upwelling of magma and a long-lasting locus of volcanic activity known as a hot spot. Some of that magma may erupt as a tremendous flood of lava. Unlike the slow, steady drifting of continents and the spreading of mid-ocean ridges, the surfacing of mantle plumes takes place in an erratic and episodic manner.

All models of the internal structure of the earth are built on inference. Even the deepest boreholes extend only about 10 kilometers. Direct and indirect studies of large igneous provinces are therefore vitally important for learning more about the nature of mantle plumes and how they may ultimately affect conditions on the earth's surface.

The most fundamental observation about these provinces is that they consist of basalt, a common, iron- and magnesium-rich rock. The large igneous features that appear in the middle of continents, where geologists can easily collect samples and determine their composition, are therefore referred to as continental flood basalts. At the edges of continents, igneous provinces are called volcanic passive margins; if they are located in the middle of the ocean, they are dubbed oceanic plateaus.

Geologists recognized the existence of continental flood basalts late in the 19th century, when they realized that several far-flung volcanic constructions actually constitute connected flows of basaltic lava. One of the most spectacular of these is the Deccan Traps in west-central India. (*Trap* is Dutch for "staircase," a reference to the steplike appearance of the eroded lavas.) Similar structures include the Columbia River basalts in the northwestern U.S.; the North Atlantic volcanic outpourings of the British Isles, the Faeroe Islands and Greenland; and the Karoo basalts in southern Africa, to name just a few. The layers of lava in a flood basalt pile up several kilometers thick. Individual flows may contain several thousand cubic kilometers of rock and may extend hundreds of kilometers.

Because of the inaccessibility of underwater igneous provinces, earth scientists only recently became aware of their resemblance to continental flood basalts. In 1981 Karl Hinz of the Feder-




**locations where seismic reflection studies seem to reveal ancient volcanism along continental margins. The largest igneous province—the Ontong Java Plateau—covers nearly two million square kilometers. Studies of these huge magmatic formations are providing new insight into the internal workings of the earth and how they may affect conditions at the surface.**

al Institute for Geosciences and Natural Resources in Hannover, Germany, suggested, based on his analysis of recordings of reflected seismic waves, that the submerged margins of many continents contain extensive, layered lava flows. Since then, investigators have obtained improved seismic data that seem to corroborate Hinz's conclusion. Igneous rocks recovered by scientific drilling projects on the continental margins off Ireland and Norway confirm that the structures seen in the seismic images do in fact have volcanic origins.

Analogous seismic studies of the Kerguelen Plateau present convincing evidence that oceanic plateaus likewise consist primarily of volcanic rocks. More recently scientists on board the *JOIDES Resolution* have recovered direct samples of rocks from two immense oceanic plateaus: Kerguelen in 1988 and Ontong Java, situated in the Pacific Ocean northeast of Australia, in 1990. The samples consist of basaltic rock, similar to that found in continental flood basalts.

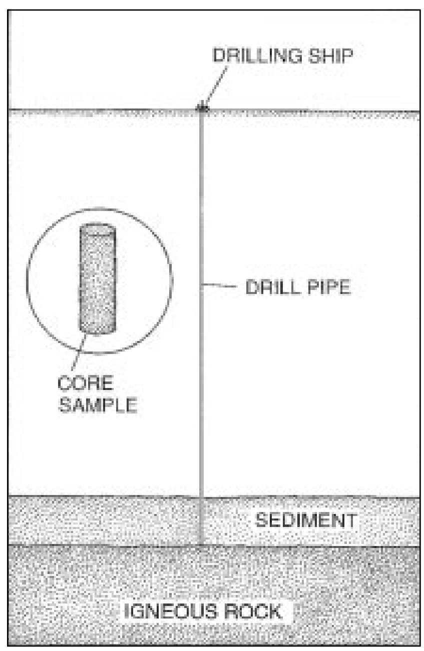
Earth scientists have wondered how such massive volcanic constructions arise. Unfortunately, the most basic information needed to answer that question—the total volumes of lava and intrusive rock in igneous provinces and the rate at which they formed—is poorly known. On land, geologists can directly measure the area of flood basalts but can only infer their depth. The crust under continents averages 35 kilometers in thickness. Only the top parts are accessible to exploration. Beneath the oceans the situation is even more difficult. Sampling just the uppermost sections of the five- to 25-kilometer-thick crust in which oceanic plateaus reside requires the services of an expensive drilling vessel.

The passage of time also disguises the true extent of ancient volcanic structures. The older igneous provinces are often heavily eroded and are therefore both disfigured and diminished in scale. John J. Mahoney of the University of Hawaii estimates that the Deccan Traps, for example, originally covered an area three times larger than at present. Moreover, in the 65 million years since the formation of the Deccan Traps, seafloor spreading has apparently dispersed sections of the original lavas to the Seychelles, the Mascarene Plateau and the Chagos-Laccadive Ridge far to the south and southwest of India.

Undersea volcanic margins and oceanic plateaus similarly have been reconfigured by the forces of change. Drill samples and seismic-reflection images of the igneous province along the Norwegian continental margin and of the Kerguelen Plateau demonstrate that when those formations first erupted, they stood above sea level and only gradually subsided to their present position in deep water. Coffin has estimated that the Kerguelen and Broken Ridge plateaus, along with most of the other oceanic plateaus and ridges in the Indian Ocean, formed above sea level and remained there for as long as 50 million years, an ample span of time in which erosion could proceed.

Despite the paucity of data and the physical transformations of large igneous provinces, we have managed to deduce the original areas covered by lava for five of the best studied of these formations: the Ontong Java and Kerguelen-Broken Ridge oceanic plateaus, the North Atlantic volcanic margins and the Deccan and Columbia River continental flood basalts. The smallest of these—the Columbia River basalts—encompasses a region larger than the state of New York;

**_JOIDES RESOLUTION_ research vessel (_left_) is the centerpiece of the Ocean Drilling Project, the most extensive international program in earth sciences. The ship incorporates a drill that penetrates the sediments that accumulate on the ocean bottom and collects core samples lying as much as two kilometers deep (_below_). In this way, earth scientists can directly analyze igneous rock from underwater continental margins and from submerged oceanic plateaus.**



the largest, the Ontong Java, is two thirds the size of Australia. To evaluate the area and volume of large igneous provinces, geologists realized that they needed to examine modern-day analogues of the gigantic igneous provinces.

No currently active volcanic regions even begin to approach the extinct large igneous provinces in magnitude. The most active modern hot spots (the big island of Hawaii and Réunion, an island lying to the east of Madagascar) cover an area of the earth's crust only about one fifth the size of the Columbia River basalts. Nevertheless, studies of the crustal structure below Hawaii contribute to the understanding of the generic forms that underlie all volcanic hot spots.

In 1982 Anthony B. Watts and Uri S. ten Brink and their co-workers at the Lamont-Doherty Earth Observatory collected seismic data by recording artificially generated sound waves that are reflected and refracted underneath the Hawaiian Islands and seamounts. The speed of those waves depends on the physical properties (such as density and elastic constants) of the rock; where those parameters change, the seismic waves will alter course. Seismic analysis reveals that the Hawaiian Islands built up as basaltic magma intruded into and collected atop the preexisting oceanic crust.

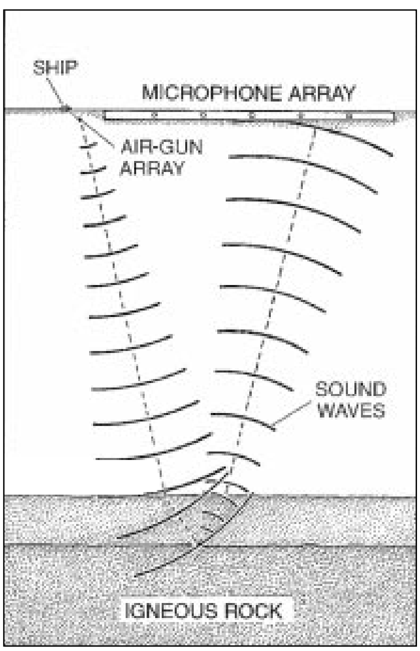
Those surface lavas evidently do not contain all of the material associated with the islands' formation, however. Below the islands and seamounts lies an anomalous zone of rock distinguished by the rapid rate at which compressional waves travel through it. Geologists think the underlying body of rock derived from the same mantle source as did the islands and seamounts. Existing techniques cannot reveal how much new igneous material was added to the previously existing oceanic crust as the islands formed. Any calculated volumes of mantle plume–derived material contained in the Hawaiian Chain therefore must be considered minimal.

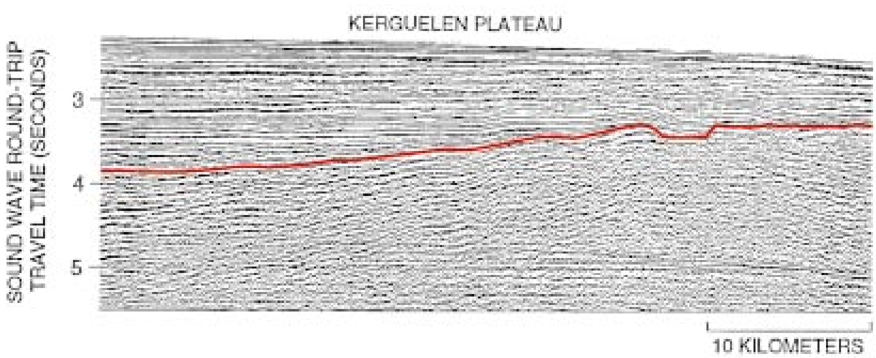
Even so, we were able to establish a basic relation between seismic-velocity structure (as well as other geophysical data) and the total volume of material encompassed within the Hawaiian Islands. We then used Hawaii as our model for inferring the distribution of volcanic material in various ancient, large igneous provinces. By our calculations, a small province such as Columbia River incorporates approximately 1.3 million cubic kilometers of rock. The giant Ontong Java province contains at least 36 million cubic kilometers of igneous rock, enough to bury the contiguous U.S. under five meters of basalt.

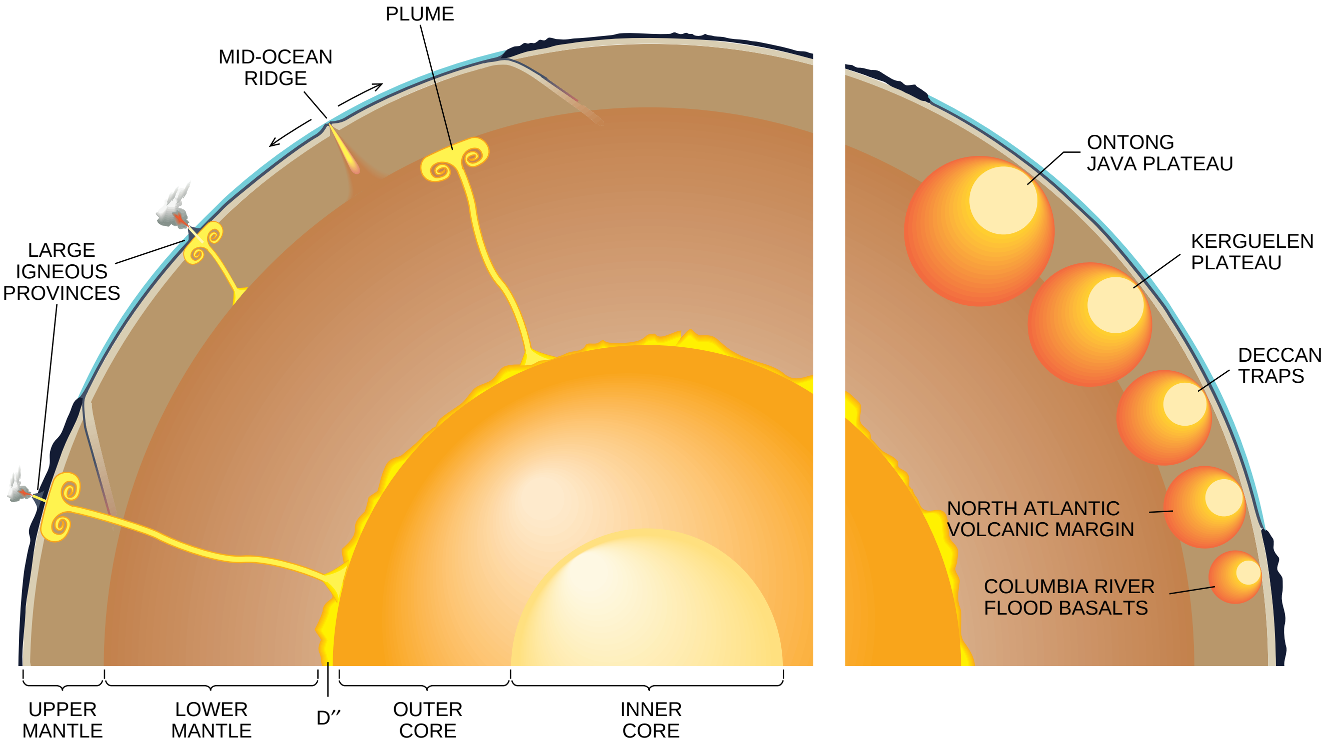
The next stage in understanding the origin and significance of the various igneous provinces is to learn how quickly they formed. Did they accumulate slowly and steadily over tens of millions of years, in much the way that new ocean crust forms at mid-ocean ridges? Or did they emerge from a series of volcanic firestorms that swiftly pumped gases and rock fragments into the air and water and that abruptly transformed the geology of large areas of the earth?

Geologists have expended a great deal of effort attempting to answer those questions for the Deccan Traps. This flood basalt has garnered special attention because it erupted some 65 million years ago, at just about the time of the major extinction at the end of the Cretaceous period, and it may have contributed to that event [see "A Volcanic Eruption," by Vincent E. Courtillot; SCI-

***RIG SEISMIC* vessel (*top right*), operated by the Australian Geological Survey Organization, collects seismic information about the ocean bottom. An air-gun array creates sound waves that reflect off the oceanic crust (*below*). Underwater microphones towed behind the ship sense the reflected waves, which are then recorded on board the ship. One such recording (*bottom right*) reveals the igneous structure (*area below red line*) underlying the Kerguelen Plateau.**






**RISING PLUMES of hot material migrate through the earth's mantle; where the head of the plume reaches the surface, a large igneous province forms (*left*). Plumes probably originate at the boundary layers between the core and mantle (the D″) and between the upper and lower mantle. The parent plumes of the most voluminous igneous provinces were so huge that they must have originated at least in part in the lower mantle, most likely at the D″. The spheres on the right depict the minimum (*white*) and maximum (*dark orange*) inferred diameters of the plumes associated with five major igneous provinces.**

ENTIFIC AMERICAN, October 1990]. Robert A. Duncan of Oregon State University, working with Vincent E. Courtillot and Didier Vandamme of the Institute of Physics of the Earth in Paris and several other colleagues, has performed radioactive dating and magnetic analysis on samples from the Deccan Traps. The scientists' results indicate that most of the lavas erupted within a span of less than one million years. Using similar methods, Ajoy K. Baksi of Louisiana State University found that the Columbia River flood basalts mostly erupted over one and a half million years.

Geologists have few direct samples on which to judge the ages and rates of formation of underwater volcanic margins and oceanic plateaus. The scant available evidence hints that submarine igneous provinces accumulated about as rapidly as those on land. Eldholm estimates that the bulk of the North Atlantic volcanic province formed in three million years or less. John A. Tarduno and his collaborators at the Scripps Institution of Oceanography, in conjunction with Mahoney, conclude that the Ontong Java province was constructed in less than three million years. Our analysis of rock dating, conducted by Hugh Davies, formerly of the Australian Geological Survey Organization, and Hubert Whitechurch, formerly of the University of Strasbourg, indicates that the Kerguelen Plateau mostly formed within a span of 4.5 million years.

From a geologic perspective, the largest igneous provinces emerged remarkably fast. In comparison, the Rocky Mountains have been rising for more than 40 million years, and the chain of Hawaiian Islands and the Emperor Seamounts have been building for at least 70 million years. Moreover, much of the volcanic activity associated with the provinces may have occurred in short, violent episodes separated by long periods of relative quiescence.

Once we had collected data on the total volume of the major igneous provinces and on how quickly they formed, we could finally deduce the magnitude of the volcanic forces that created them. The eruptions that built up the Ontong Java province unleashed between 12 and 15 cubic kilometers of igneous rock each year; Deccan volcanism produced between two and eight cubic kilometers annually. Assuming that the creation of igneous provinces, like other eruptive processes, occurs in fits and starts, the pace of crustal production may have been far greater in some years. For reference, Roger L. Larson of the University of Rhode Island estimates that the global network of mid-ocean ridges has yielded between 16 and 26 cubic kilometers of ocean crust a year over the past 150 million years. In other words, individual igneous provinces have generated new crust at rates comparable to or greater than that of the global seafloor spreading system.

Large igneous provinces build up so swiftly compared with the churning of the earth's deep interior that they must have derived from single, discrete sources. If one knows the volume of basaltic rock contained in those provinces, one can evaluate the dimensions of the hot plumes in the mantle that led to their genesis. Only a fraction of the plume material actually melts and reaches the crust, and this fraction presumably is smaller at greater depths, where increased pressures tend to keep the mantle rocks in a solid state. Hence, less melting should take place under thick continental lithosphere than below thin oceanic lithosphere. Stephen M. Eggins of the Australian National University and Shen-su Sun of the Australian Geological Survey Organization estimate that the molten, basaltic portion of the plume (the part that gives rise to the surface volcanism and to subsurface igneous intrusions) accounts for from 5 to 30 percent of the plume's total volume.

We used those numbers to calculate diameters for the thermal anomalies in the mantle associated with the five best-studied igneous provinces. For simplicity, we assumed the rising plumes to be spherical because such a shape represents the most efficient way to transport material and so permits the plume to move at a plausibly slow velocity through the earth's interior. Our analysis suggests that the Ontong Java province must be derived from a mantle plume at least 600 kilometers, and possibly as much as 1,400 kilometers, in diameter.

That size carries special significance for geologists. It indicates that large plumes must contain at least some material from the lower mantle, more than 670 kilometers below the surface. At that depth, the velocity of seismic waves changes abruptly, probably because of a change in the mineral structure of the mantle rocks. The issue of whether the whole mantle mixes or whether the upper and lower parts of the mantle behave as independent systems that circulate separately strongly divides geophysicists.

Our work favors models that allow for at least some interaction between the upper and lower mantles. In our view, the largest plumes originate in the lower mantle, most likely at the D″, a region having unusual seismic properties that lies just outside the core. Smaller plumes may arise at the 670-kilometer-deep transition zone between the upper and lower mantle.

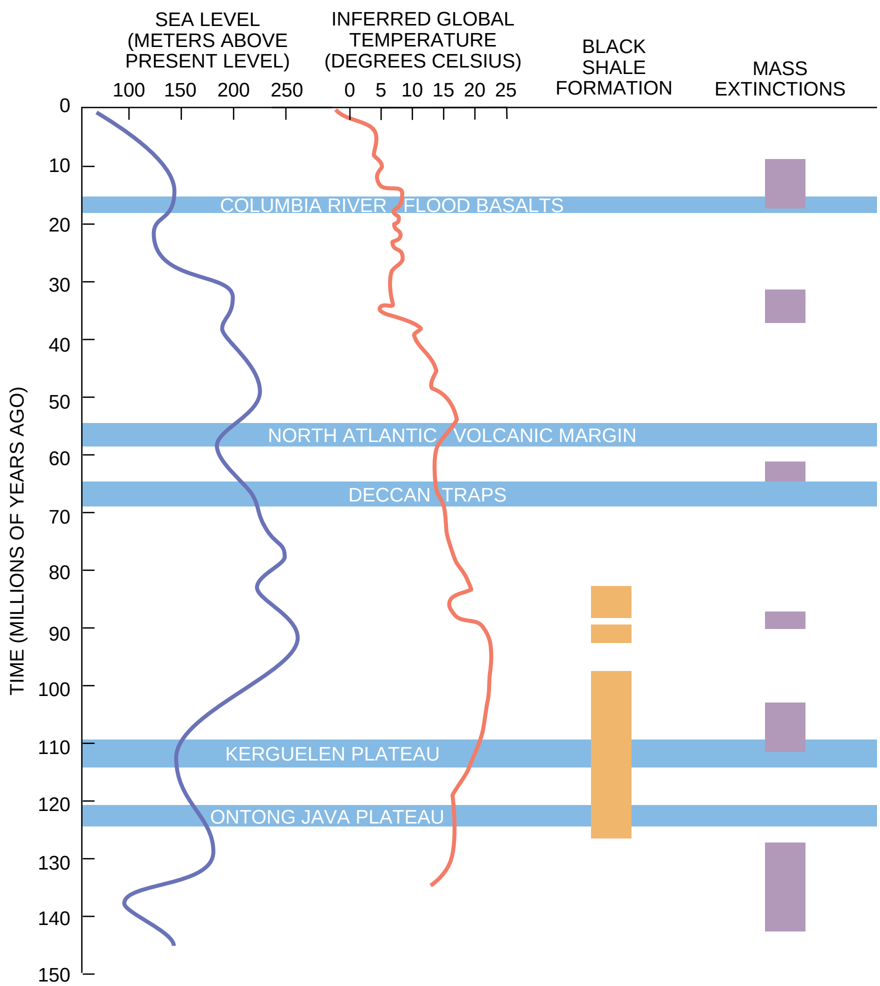
Regardless of where mantle plumes originate, their attributes and effects at the surface depend strongly on the temperature, composition and physical state of the material they encounter just beneath the lithosphere. These factors, combined with the local strength of the lithosphere, determine the volume, timing and position of the eruption at the surface. When plumes rise under continental masses, they may help force the continents apart at a weak spot and then induce the formation of extensive volcanic features along the margins of the rifted landmass [see "Volcanism at Rifts," by Robert S. White and Dan P. McKenzie; SCIENTIFIC AMERICAN, July 1989]. Under certain circumstances, the plume may penetrate the thick central regions of continental blocks and give birth to a continental flood basalt.

If the upwelling mantle plume surfaces underneath the seafloor, it may give rise to an oceanic plateau. Laboratory experiments suggest that a drawn-out tail of hot material should lag behind the spherical head of the plume, yielding a long-lived, focused source of magma. Over millions of years, plate motions cause the ocean floor to migrate over the site of the hot spot. Where lava erupts at the surface, it gradually constructs a linear submarine ridge or a sequence of islands and seamounts. The Hawaiian-Emperor chain presumably developed in this manner, although how a plume could persist for more than 70 million years remains a puzzle.

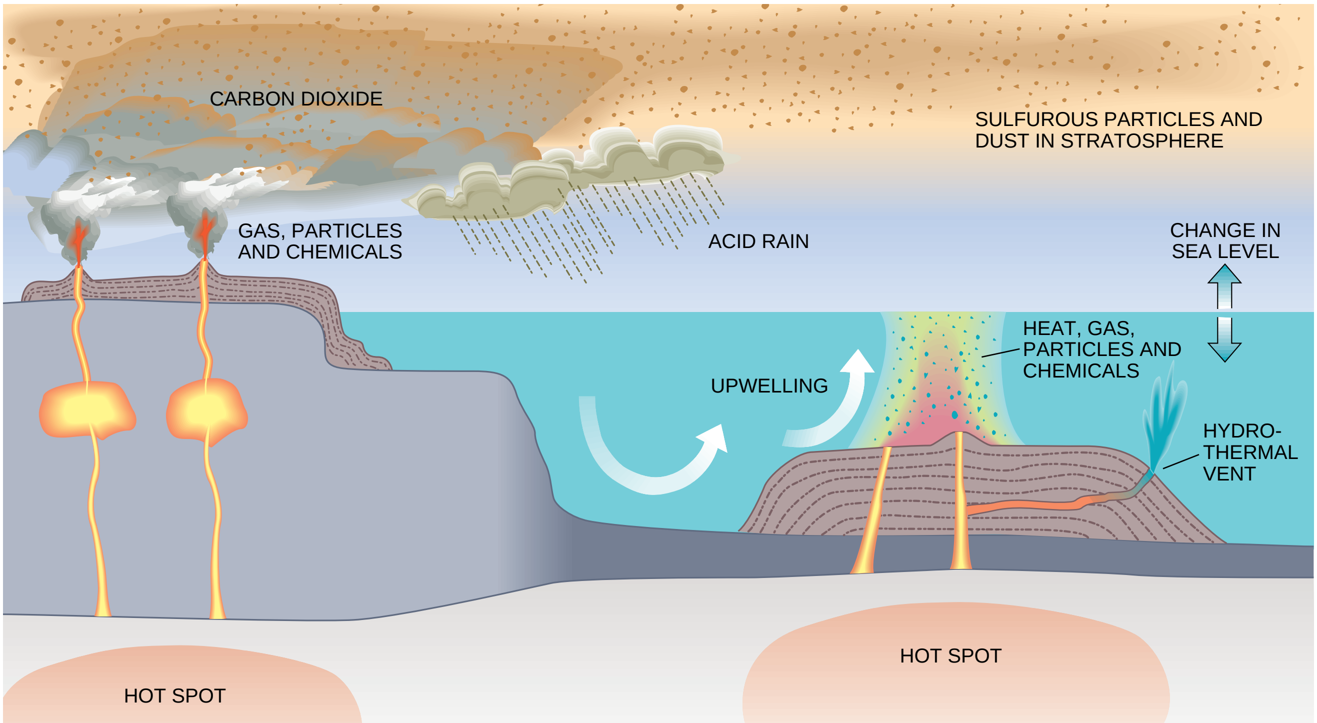
As the 1991 eruption of Mount Pinatubo in the Philippines documents, even moderate volcanic outbursts can severely damage the local environment. Yet such geologic events are incidental twitches compared with the convulsions of magmatic activity during the formation of large igneous provinces. One can thus begin to imagine that those ancient eruptions had profound consequences. In 1972 Peter R. Vogt of the Naval Research Laboratory first proposed that the surfacing of a mantle plume would lead to physical and chemical changes around the globe; the environmental effects associated with these changes could strongly influence the evolution of life.

The current fascination with the greenhouse effect and global change has renewed interest in Vogt's ideas. Stephen Self of the University of Hawaii and Michael R. Rampino of New York University note that the environmental impact of a large igneous province depends in part on whether it forms on land or underwater. Ocean plateaus and volcanic passive margins modify the geometry of the ocean basins and alter the global sea level. We estimate that accumulation of the Ontong Java plateau material elevated sea level by about 10 meters. Oceanic plateaus and volcanic margins may block or constrict circulation in ways that influence large-scale circulation, erosion and sedimentation, especially if the volcanic activity occurs in a sensitive location between ocean basins.

Because they are much denser and




**ENVIRONMENTAL EFFECTS from the eruption of large igneous provinces include fluctuations in the global sea level. Powerful volcanic activity could alter the chemistry and circulation of the atmosphere and oceans, influencing the evolution of life. Volcanoes release carbon dioxide, which may contribute to greenhouse warming; warmer temperatures promote the formation of black shale, in part by enhancing global biomass production. In contrast, several mass extinctions of the past 150 million years coincide with the appearance of igneous provinces, hinting at a causal relationship.**




**PHYSICAL AND CHEMICAL EFFECTS accompany the appearance of large igneous provinces. Volcanoes emit carbon dioxide, which can elevate the global temperature. Sulfurous particles and dust in the stratosphere could induce acid rain and block out sunlight. Heavy metals and other chemicals emitted during eruptions would alter the composition of the land, air and water. Topographic changes associated with extensive volcanism would modify oceanic circulation and change sea level; heat and hydrothermal activity cause upwelling, further transforming underwater conditions.**

more massive than the atmosphere, the oceans are far more able to absorb and dilute gases and heat, a factor that tends to ameliorate the consequences of volcanic eruptions. On the other hand, submarine volcanism and associated hydrothermal activity may leach trace metals such as arsenic, which are poisonous to marine life. Heat produced by submarine eruptions induces ocean-bottom waters to rise to the top, altering the circulation of the surface waters and disrupting the organisms living there.

Carbon dioxide emitted at an undersea volcanic site spreads through the ocean. In so doing, the compound could increase the alkalinity of the seawater, thereby affecting both marine life and climate. If elevated levels of carbon dioxide lead to global warming, ocean circulation should turn sluggish, in which case it would carry less dissolved oxygen. Oxygen-deprived waters may contribute to the formation of carbon-rich black shale, as documented by the late Seymour O. Schlanger of Northwestern University, Michael A. Arthur of Pennsylvania State University and Hugh C. Jenkyns of the University of Oxford.

The atmosphere must have absorbed an extensive outpouring of carbon dioxide around 120 million years ago, when the tempo of volcanism peaked. Larson speculates that a massive thermal instability in the D″ layer initiated so-called superplumes that ultimately supplied the Ontong Java and Kerguelen oceanic plateaus, along with several other smaller plateaus in the Pacific Ocean. Kenneth G. Caldeira of Penn State and Rampino have run computer models of the geochemical cycle based on their assumptions about the quantity of carbon dioxide in the atmosphere during that era. From the models, the workers infer that worldwide temperatures averaged from 7.6 to 12.5 degrees Celsius higher than today's mean, if one takes into account the different geography and higher sea level at the time.

Where volcanic eruptions occur on dry land, they directly alter the physics and chemistry of the atmosphere. Alan R. Huffman of Exxon Exploration Company in Houston calculates that a single flood basalt event that generates 1,000 cubic kilometers of lava (the volume of a typical flow in the Columbia River province) emits 16 trillion kilograms of carbon dioxide, three trillion kilograms of sulfur and 30 billion kilograms of halogens. The thousands of such episodes that must occur in the accumulation of an individual large igneous province would modify the atmosphere in ways that would dwarf the effects of modern, human-generated pollutants.

Explosive eruptions of silica-rich rock often carry sulfurous particles into the stratosphere, where they are converted to tiny droplets of sulfuric acid, according to Rampino and Self. Basaltic lavas emit about 10 times as much sulfur per unit volume as do silica-rich lavas; Charles B. Officer of Dartmouth College and his co-workers deduce that if the gases and particles produced during basaltic eruptions were shot into the stratosphere, they could cause short-term plagues of acid rain, worldwide darkening and global cooling. Richard B. Stothers of the National Aeronautics and Space Administration Goddard Space Flight Center, along with several others, hypothesizes that massive lava fountains and vigorous atmospheric convection taking place over eruptive vents in continental flood basalts could supply the sufficient upward momentum to inject material into the stratosphere.

The powerful Lakagigar eruption in Iceland in 1783–84 illustrates the potential disruptive effects of flood basalt volcanism. Although the Lakagigar eruption poured out only about 15 cubic kilometers of lava, local temperatures declined noticeably in the following years. About three quarters of all livestock in Iceland died, probably the result both of the deterioration of the climate and of the emissions of acid gas;

the resulting famine killed about one fourth of Iceland's human population. Dust veils, fog and haze appeared over most of Europe and adjacent parts of Asia and Africa for many months after the eruption.

Scientists are uncovering clear evidence that the environmental consequences of flood basalt volcanism have in fact contributed to mass extinctions. The most severe extinction in terrestrial history occurred 248 million years ago, when the Siberian Traps formed. At that time, roughly 95 percent of all marine species perished; the ensuing evolutionary free-for-all marked the first appearance of the dinosaurs.

The biological repercussions of the eruption of large igneous provinces may depend in part on the state of the global environment at the time. When the environment is already stressed by other factors, such volcanism may trigger rapid climatic, oceanographic and biotic changes. When the environment is robust, few effects may show up in the geologic record.

Oddly enough, the most voluminous igneous province, the Ontong Java, produced virtually no detectable extinctions. On the contrary, its formation coincides with the deposition of black shale, which is suggestive of an interval of enhanced biological activity. We propose that the deleterious effects associated with the Ontong Java were minimal in part because it erupted underwater, as indicated by Ocean Drilling Program studies led by Loren Kroenke of the University of Hawaii and Wolfgang H. Berger of the Scripps Institution of Oceanography. We also suspect that the global environment was in a resilient state 120 million years ago, although such an assertion is difficult to quantify and prove. In contrast, the emplacement of the smaller Kerguelen and Broken Ridge provinces, which took place about 110 million years ago, coincided with a mass extinction as well as with major deposits of black shale.

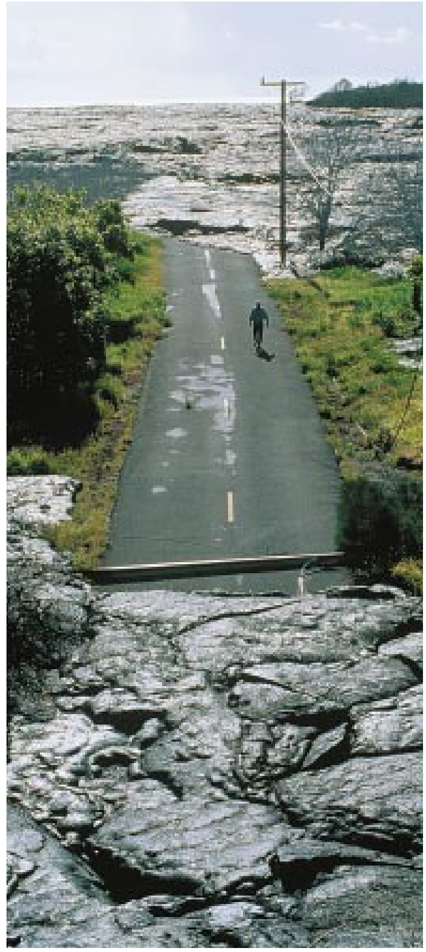
The eruptions that built up the Deccan Traps, and simultaneous volcanism along the margins of western India and along the Seychelles and the Mascarene Plateau, took place approximately 65 million years ago—just about the time that nearly half of all species of life, including the dinosaurs, went extinct. Scientists continue to debate whether to attribute those changes to the impact of a sizable asteroid or to more earthly explanations. Even if an asteroid was the primary agent of the extinction, the Deccan eruptions may have contributed to an environmental deterioration that could have magnified the repercussions of the impact.

Another significant but less celebrated change in the global environment took place 10 million years after the demise of the dinosaurs, during the emplacement of the North Atlantic volcanic margins. At that time, many deep-ocean foraminifera and land mammals became extinct, and hydrothermal activity was high. Moreover, David K. Rea and his co-workers at the University of Michigan, along with Ellen Thomas of Yale University, find evidence of major transformations in deep water and in atmospheric circulation.

Analysis of the oxygen isotopes taken in by foraminifera indicates that ocean temperatures 55 million years ago, during the early Eocene epoch, were warmer than at any other time during the past 70 million years. Eldholm, working with Thomas, recently postulated that the balmy ocean surface of the early Eocene could have resulted from the outgassing of carbon dioxide during the eruptions along the North Atlantic volcanic margins. Ash layers dating from 55 million years ago cover large regions of northwestern Europe, affirming the violence of the volcanism. Higher atmospheric temperatures could have led to the formation of a layer of warm water atop the oceans at high latitudes. Such a warm surface layer would tend to resist mixing with underlying, cooler waters. The consequent changes in deepwater circulation could have been fatal for many ocean-bottom species.

Even the relatively small Columbia River flood basalts coincided with a mass extinction 16 million years ago. At about that time, the earth began to experience an ongoing cycle of ice ages, as Maureen E. Raymo of the Massachusetts Institute of Technology points out. Perhaps the global environment was already so frail that even a moderate eruption could have had a substantial impact.

Obviously, geologists have—literally and figuratively—only just scratched the surface of large igneous provinces. Current knowledge of these formations amply demonstrates that they contain crucial information about the internal workings of the earth and about the natural causes of global change. Earth scientists are now intensifying their efforts to produce better seismic images, to undertake more field and laboratory studies, to perform additional modeling and to conduct more scientific drilling. Those labors promise even better insight into the intimate links between the earth's inner and outer realms.



**LAVA FLOW from Kilauea Volcano in Hawaii only hints at the magnitude and potential environmental impact of the flows incorporated into large igneous provinces. Individual sheets of lava in the Columbia River flood basalts, for example, extend for hundreds of kilometers and stand several meters thick.**

FURTHER READING


CONTINENTAL FLOOD BASALTS. Edited by J. D. Macdougall. Kluwer Academic Publishers, 1988.

HOTSPOTS, MANTLE PLUMES, FLOOD BASALTS, AND TRUE POLAR WANDER. R. A. Duncan and M. A. Richards in *Reviews of Geophysics,* Vol. 29, No. 1, pages 31–50; February 1991.

GEOLOGICAL CONSEQUENCES OF SUPERPLUMES. Roger L. Larson in *Geology,* Vol. 19, No. 10, pages 963–966; October 1991.

MAGMATISM AND THE CAUSES OF CONTINENTAL BREAK-UP. Edited by B. C. Storey, T. Alabaster and R. J. Pankhurst. Special Publication, No. 68. Geological Society of London, 1992.

LARGE IGNEOUS PROVINCES: CRUSTAL STRUCTURE, DIMENSIONS, AND EXTERNAL CONSEQUENCES. M. F. Coffin and O. Eldholm in *Reviews of Geophysics* (in press).

# Evolutionarily Mobile Modules in Proteins

*Many proteins consist of a fairly small set of modular elements. How these units spread and multiplied during evolution is not altogether clear, but a pattern may be emerging*

by Russell F. Doolittle and Peer Bork

Molecular biologists and biochemists have learned in recent decades that many proteins consist of domains, or discrete blocks of amino acids. Many of these domains have well-defined functions that contribute to the overall activity of a protein. Furthermore, some of these modular units frequently move about within and between proteins during evolution. The evolutionary mobility of these modules is not restricted to hops within the genetic material of a single species: in some cases, the modules are apparently able to travel laterally across species lines—even moving from animal cells to bacteria, for example.

Because a compartmental delineation has also been found in the regions of genes that encode proteins, many biologists have been convinced that these structural features are reflections of the same underlying phenomenon. They believe that each genetic coding region corresponds to a specific structural feature in a protein. We and our colleagues take a somewhat different point of view. The weight of evidence, in our opinion, falls to the argument that the subdivision of genes into separate coding parts is a far more recent development by evolutionary standards.

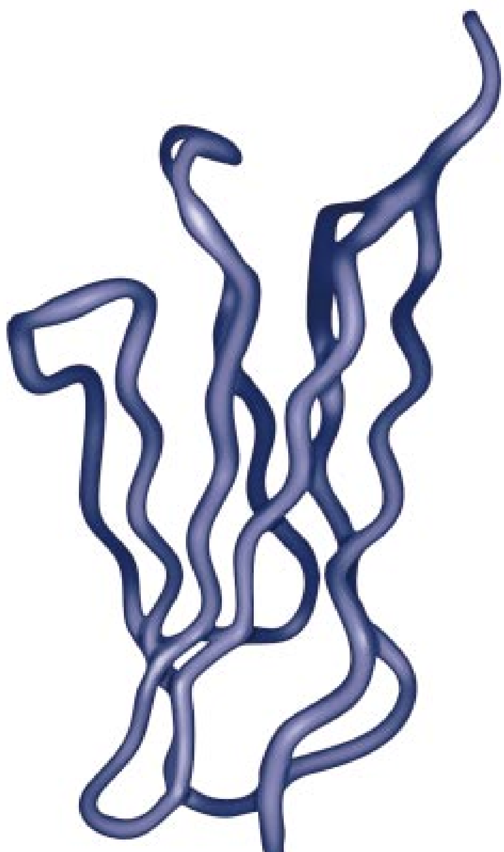
Proteins are long chains of small molecules called amino acids. Twenty different kinds of amino acids, each with its own shape and chemical character, make up all the proteins found in nature. All the properties of a protein depend on which of the 20 amino acids are used in its construction and, particularly, on the order in which they are strung together. Most notably, the amino acid sequence determines how the protein will fold up into an active, three-dimensional body.

The length of a sequence plays an important part in that determination. Chains of amino acids—usually called polypeptides—may be up to many thousands of amino acid units long. (The record to date is for titin, a muscle protein that contains upward of 30,000 amino acid residues.) Short chains, however, are not large enough to have sufficient intramolecular attachments to lock themselves into a single conformation; they tend to flop from one form to another. Ordinarily, when a polypeptide chain contains 30 or 40 residues, it begins to have enough internally cohesive forces to give it a predominant shape, although it may still need additional stabilization from bound metal ions or disulfide bonds between pairs of the amino acid cysteine.

In a constant environment, any protein containing more than some minimum number of amino acid units will always fold itself in the same way. That environment may be the dilute salt solution that constitutes many biological fluids or the greasy confines of a biological membrane; it can also include nearby proteins or even other parts of the same long polypeptide chain.

A sequence that folds spontaneously into a characteristic shape under a defined set of circumstances is called a domain, but that formal definition is seldom applied with rigor. More often than not, the term is applied to any part of a protein that can be defined as structurally distinct from the rest. Some small proteins are completely embodied in a single domain; many others are composed of two or more domains; and some are made up of many domains, the shapes of which may be very similar or very different.

RUSSELL F. DOOLITTLE and PEER BORK analyze the structures of proteins to gain insights into evolution. Doolittle is professor of biology at the Center for Molecular Genetics of the University of California, San Diego. He began his studies of protein evolution while he was a graduate student in biochemistry at Harvard University in the late 1950s. Doolittle is a member of the National Academy of Sciences and other scientific associations. This article is his third for *Scientific American*. Bork is a project leader at the Max Delbrück Center for Molecular Medicine in Berlin and a visiting scientist at the European Molecular Biology Laboratory in Heidelberg. He also holds a teaching appointment at Humboldt University in Berlin.



CD2 DOMAIN 2

The most straightforward way to identify a domain is to determine its structure through x-ray diffraction of protein crystals or nuclear magnetic resonance (NMR) studies. Once the amino acid sequence of a domain has been identified, researchers can find other related domains without recourse to structural studies: they can simply look for amino acid sequences that are similar to those of familiar structures. That shortcut is extremely helpful because many more amino acid sequences are known than are structures from x-ray diffraction or NMR. Even in the absence of experimental determinations, it is often possible to infer the existence of a domain on the basis of sequence alone. By looking at the structural and sequence similarities of proteins in this way, one can learn much about their evolution.

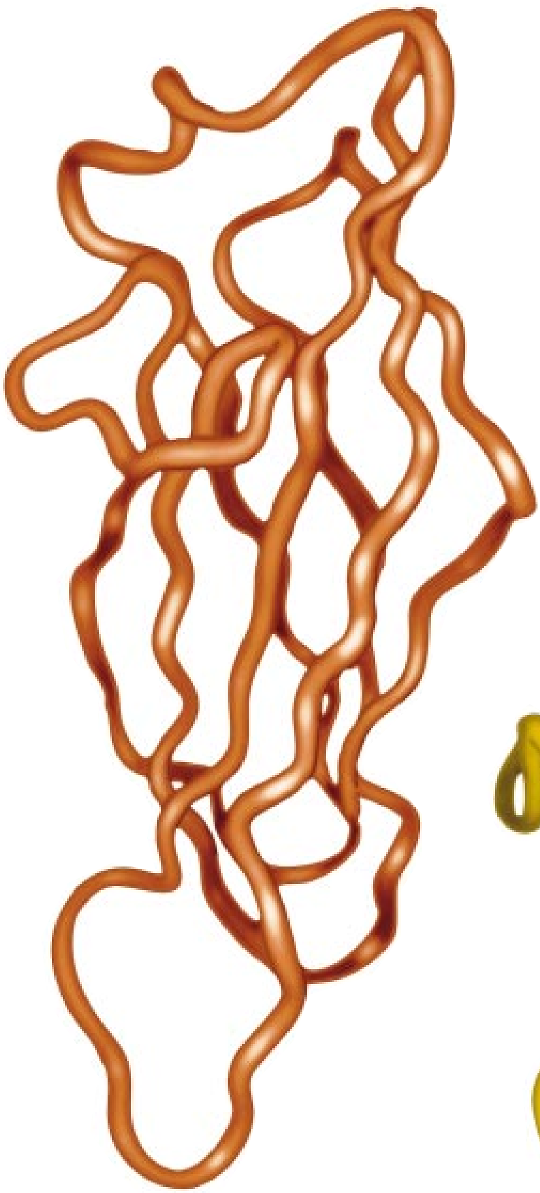
Until the early 1970s, the conventional wisdom about how proteins evolved centered mostly on "duplication and modification." The gene for a given protein is occasionally duplicated by various recombination processes during which genetic information is exchanged between strands of DNA. Sometimes the duplication results in a second gene that can then undergo further modification or mutation to produce a new protein with a novel function. Alternatively, the duplicated DNA can be in tandem, in which case the original protein is elongated and may assume novel properties as a result. As comparisons of amino acid sequences make obvious, these kinds of internal duplications have clearly given rise to many extended proteins, ranging from small ones, such as the bacterial ferredoxins that have only 56 amino acid residues, to large ones, such as bacterial beta-galactosidase, which has more than 1,000.

A hidden aspect of protein evolution came to light about 20 years ago, however, when Michael G. Rossmann of Purdue University determined the three-dimensional structure of the enzyme lactate dehydrogenase by x-ray diffraction. One part of that molecule, he noticed, closely resembled features of some other proteins he had seen. Specifically, a part of the enzyme that bound to a cofactor had obvious counterparts in other dehydrogenases.

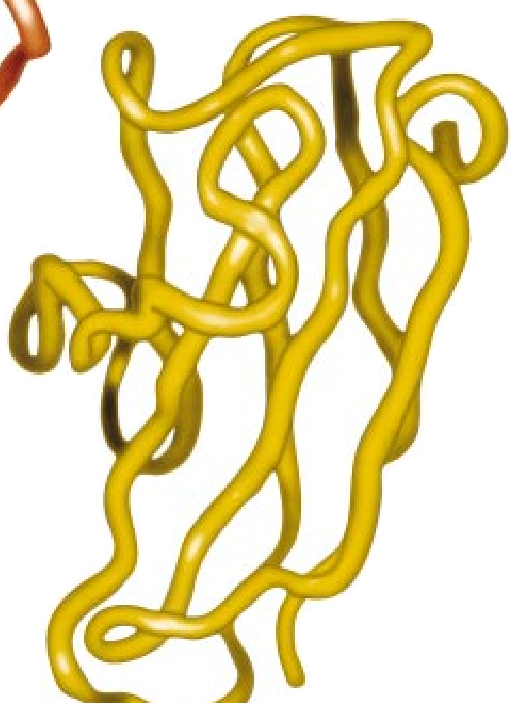
What made those structural similarities noteworthy was that they did not always occur in the same relative parts of the molecules. It seemed as though during evolution the unit had moved around within the linear amino acid sequence without losing its function of binding to a cofactor. Rossmann suggested that proteins were constructed of modules—what we would now consider domains—that had appeared early in life's history and had been assembled in different combinations.

His observation presented a possibility for protein evolution that greatly aug-

**GENETICALLY MOBILE MODULES have been found in many proteins. Two types of these modules, or domains, are shown here. The Fn3 and the GHR domains are examples of fibronectin type III modules. The PapD and CD2 domains are immunoglobulin domains. These modules are linear sequences of amino acids that can fold themselves into consistent, recognizable structures with specific biochemical properties. During evolution, these domains can move as discrete units from one protein to another, which helps new types of proteins to appear.**

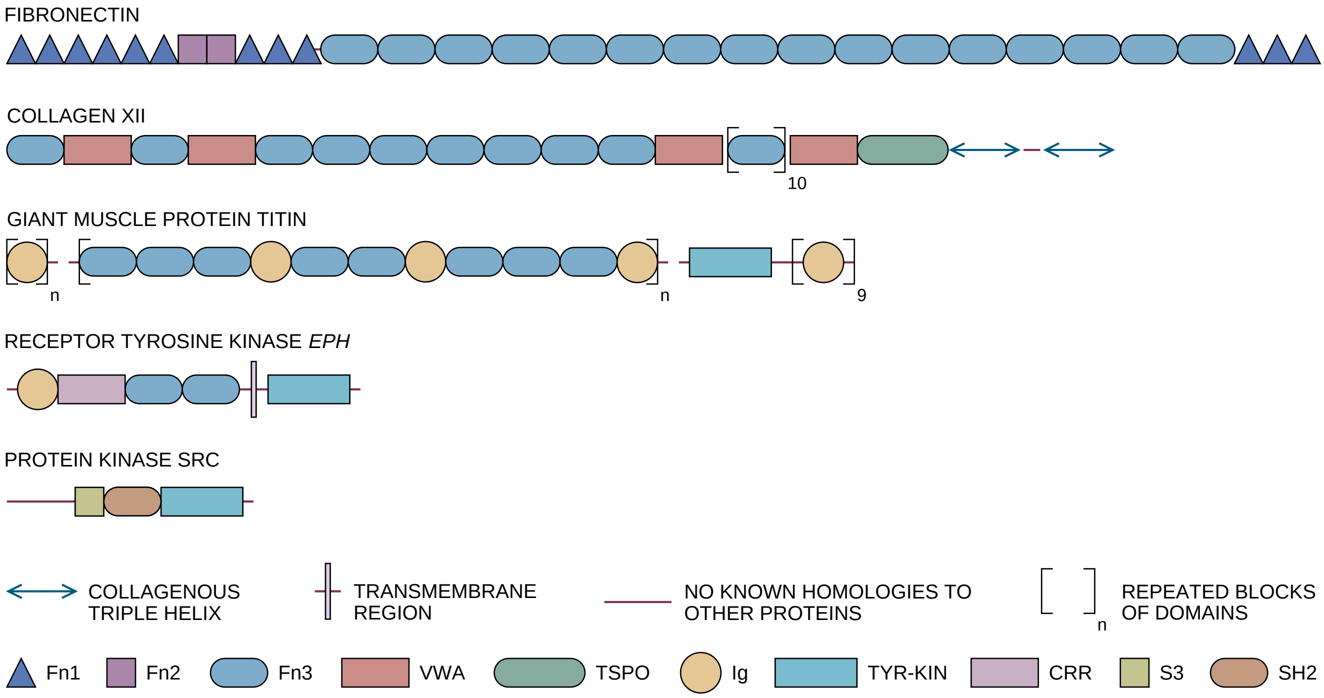


**PapD DOMAIN 1**



**GHR DOMAIN 1**

**Fn3 (10TH DOMAIN)**




**LIKE BEADS IN A NECKLACE, domains appear as distinct subunits of modular proteins. Some proteins, such as fibronectin, collagen XII and the muscle protein titin, contain many repeats of the same few domains.**

mented what could be accomplished purely through duplication and modification. If new proteins could be created by recombining the components of other ones, then protein diversity could grow explosively.

Rossmann's suggestion proved prescient. The amino acid sequences for numerous large proteins have been determined, and many have the kind of grossly repetitive structure that one might expect of a string of mobile modules. For instance, the protein fibronectin is made up of two long chains, each of which contains more than 2,000 amino acid residues. Casual inspection revealed that the chains of this large extracellular protein consist of several runs of three different types of repeated sequences. The repeats, which are referred to as Fn1, Fn2 and Fn3, have lengths on the order of 45, 60 and 100 amino acids, respectively. (The repeats are "imperfect," which means that all the repeats of a given type are not exactly identical.) Presumably, each type of repeat can fold up independently as a true domain, and the overall protein must be like a long necklace made from three types of beads.

Surprisingly, sequences similar to Fn1, Fn2 and Fn3 were subsequently observed in a huge assortment of other animal proteins. Much the same is also true for a number of other identified domains. The small protein called epidermal growth factor consists of a single domain (53 amino acids long in the human version) that is tightly folded and pinned together by three disulfide bonds. Similar domains have been identified by their sequences in more than 100 proteins, in which they occur from one to more than 30 copies.

The functions of many of these modules are not altogether clear, but many of them do bind to or recognize particular substances. There is a family of lectins that bind to various carbohydrates. Similarly, the immunoglobulin domain, a feature of antibodies and other molecules in the immune system, is well known for its binding capabilities. Some domains may serve as recognition tags that identify a protein as "belonging" to a certain tissue. Many appear to be mere linkers or spacers, innocuous connecting units. Finally, some may have no function at all. It appears, therefore, that many domains can and do move within and between proteins during evolution. As long as no damage or loss of function results from a shuffle, the evolutionary cost of maintaining a domain in a new setting is insignificant. That result is a corollary of the theory of neutral evolution, but it might be phrased as a rule borrowed from professional basketball: "No harm, no foul."

When Rossmann first suggested that modular units might move within and between proteins, no one pondered very seriously the genetic mechanisms that might account for those rearrangements. Not long afterward, however, molecular biologists happened on an unexpected feature of genes that seemed to offer an explanation. As James Watson and Francis Crick had learned in 1953, genetic information is inscribed in the double helices of deoxyribonucleic acid (DNA) molecules. Just as proteins are strings of amino acids, DNA molecules are strings of nucleotide bases. The DNA nucleotide sequences are copied, or transcribed, as complementary strands of messenger ribonucleic acid (RNA). Inside cells, miniature factories called ribosomes assemble proteins from the instructions in the RNA: each codon, or three-base sequence in the RNA, corresponds to an amino acid.

The surprising observation made in the mid-1970s was that the DNA coding for a polypeptide can be interrupted by noncoding sequences—arrays of bases that do not correspond to the sequence of amino acids found in the ultimate protein product. The noncoding sequences are excised by a splicing mechanism before the messenger RNA strand is translated into a polypeptide.

The discovery of those interruptions in genes prompted Walter Gilbert of Harvard University to suggest that the noncoding sequences, which he named introns (for intervening sequences), facilitated the exchange of the coding parts of the genes, which he termed exons (for expressed sequences). The thought was that the additional distance between coding segments would lead to proportionately more opportunities for recombinations, which depend on random breaks in the DNA. It was suggested that similarities between the sequences of introns might promote misalignments and unequal crossovers of DNA during recombination, which would make the rearrangement of genes easier. Although there was no basis at that time for supposing that introns would have similar sequences, subsequent research has shown that introns are a haven for large numbers of mobile genetic elements. The similar sequences in those mobile elements can contribute to genetic delinquencies during meiosis, the cell division process that gives rise to eggs and sperm.

Of course, many creatures do not engage in meiosis; have those organisms missed out on a splendid way to assemble new proteins? Introns that interrupt protein coding are found only in the DNA of eukaryotes, organisms with discrete nuclei. The genes for bacterial proteins do not contain introns: every set of three bases corresponds to an amino acid in the protein. (A few types of introns that do not interrupt proteins have been found in bacteria, but they are not properly part of this discussion.)

The absence of introns from bacterial protein genes led Ford Doolittle of Dalhousie University and James Darnell of the Rockefeller University to suggest independently that bacteria may have possessed introns in the past but had lost them. Presumably, their genomes had been streamlined during evolution to make their replication more efficient. In short, the introns had been around since the beginning of life; it was the short coding sequences that were engendered separately.

Doolittle and Darnell's proposal led to a still unsettled debate about whether introns appeared "early" and are fundamental to the origin of all proteins or whether they came "late." The latter proposition was advanced by Thomas Cavalier-Smith, now at the University of British Columbia, who theorized that introns might be invasive bits of nucleic acids, referred to as transposable elements, that originated in the symbiotic organisms that eventually became the mitochondria and other organelles of eukaryotic cells. His idea has been extended by Donal Hickey of Ottawa University, among others.

As it turns out, the stretches of DNA that encode the evolutionarily mobile modules in proteins are frequently, but not always, flanked by introns. In other words, the structural units in many proteins are encoded by exons. That observation fostered a widespread belief that all exons are evolutionarily mobile and correspond to potential modular building blocks in proteins.

In our view, this notion is mistaken on two counts. One is that, as Lázló Patthy of the Institute of Enzymology in Budapest first pointed out, all exons can be shuffled, but only a fraction of such shuffled units will be genetically compatible—that is, many of them cannot be sensibly translated in some new positions. When an intron falls in a coding sequence, it will occupy one of three types of positions: squarely between two codons (type 0), between the first and second positions of a codon (type 1) or between the second and third positions (type 2). If that intron and the coding sequence adjacent to it are shuffled into a new location, the intron must adopt the same type of position—otherwise the shifted codons will be translated improperly and a nonsensical amino acid sequence will result. If introns were distributed randomly, we would expect to find that after a shuffle, only one third of the new exon combinations would be in phase. Curiously, the overwhelming majority of the genes for the most frequently shuffled modules are flanked by introns that are of type 1.

The other fundamental reason why only some exons are evolutionarily mobile is that only true domains—those that can fold completely and independently—will be able to survive in new protein settings. Smaller, less self-sufficient sequences would be unable to fold and would lose their identity. Moreover, if a shuffled unit were to land between two exons that were not themselves true domains, then the product of the gene receiving the addition might not be able to fold itself properly either.

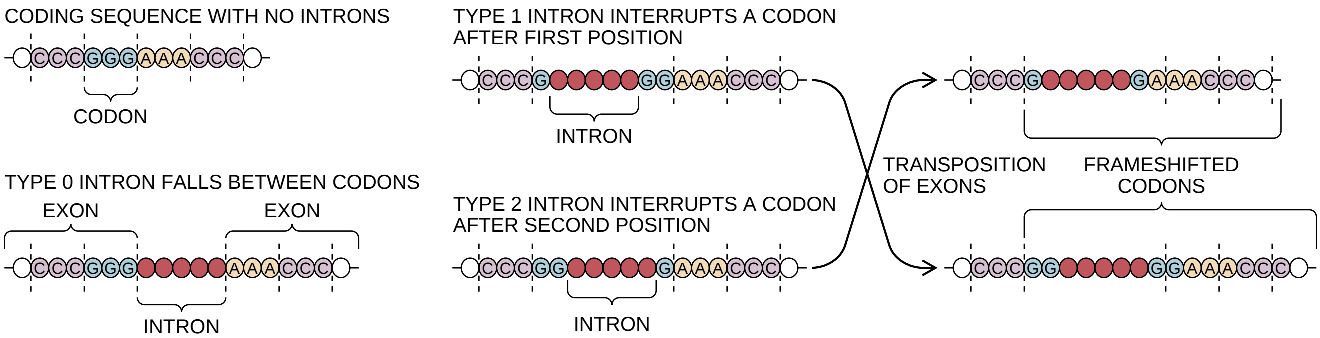
These two factors, one genetic and the other structural, contribute greatly to why mobile domains are so often found in one another's company. Not infrequently, proteins that contain one mobile domain contain others; some proteins are mosaics that contain as many as five different commonly shuffled modules. These types of proteins are both genetically and structurally tolerant of the shuffling process.

The observation that many modules are encoded by exons has been interpreted as support for the idea that the primeval organisms assembled all their proteins from an inventory of exon-encoded primitive structural components. Several points argue against such an interpretation, however. For one, simple arithmetic proves that the hypothetical early exons would have been too small to produce protein components that

## Sizes of Some Mobile Domains

Evolutionarily mobile domains vary in size. Some hold their shape in part because of disulfide bonds between pairs of their cysteine residues. Other domains are stable without such bonds.

| DOMAINS CONTAINING DISULFIDE BONDS | APPROXIMATE NUMBER OF AMINO ACID UNITS | NUMBER OF CYSTEINES |
|---|---|---|
| Somatomedin B | 40 | 8 |
| Complement C9 | 40 | 6 |
| EGF | 45 | 6 |
| Fn1 | 45 | 4 |
| Fn2 | 60 | 4 |
| "Apple" (Sushi) (ßGP26) | 61 | 4 |
| Ovomucoid | 68 | 6 |
| VWF-C | 72 | 8 |
| Kringle | 80 | 6 |
| Kunitz | 80 | 6 |
| Link | ~100 | 4 |
| Scavenger receptor | 110 | 6 |
| Fibrinogen-related domain | 250 | 4 or 6 |
| **NO DISULFIDE BONDS** | | |
| Collagen | 18 | |
| Leucine-rich | 25 | |
| "Gla" | 35 | |
| Collagen-binding | 50 | |
| Lectinlike | 100 | |
| Fn3 | ~100 | |
| SH2 | 100 | |
| SH3 | 100 | |




**INTERRUPTED STRUCTURE of some genes that encode proteins divides them into exons and introns (expressed and intervening sequences). Triplets of bases in exons are translated as amino acids in proteins. Introns can interrupt coding sequences in three different positions. If exons and introns moved randomly, transpositions between different types of introns would cause frameshift mutations. That problem counts against the idea that all exons encode movable modules.**

could fold on their own. The average size of the known exons in vertebrate genomes today is 135 nucleotide bases, which corresponds to a polypeptide of a mere 45 amino acids. A sequence that short usually needs auxiliary stabilization to fold into a stable conformation.

Keep in mind, too, that the proponents of the "intron early" theory contend that introns are constantly being lost over time. They have been forced to that conclusion by the sporadic occurrence of introns in different species. This inconstant pattern could be the result of either gain or loss of introns, but if one is wedded to the notion that introns were there from the beginning, then the only explanation is loss. Accordingly, the earliest exons would have been even smaller, encoding polypeptides that could not reasonably be expected to fold into domains on their own.

Another argument against the use of modern mosaic proteins as examples of early building blocks concerns the distribution of domains among proteins. Far and away, the majority of mobile modules known to date are found exclusively in animal proteins. At this point, we do not really know when or where most of them arose initially. Perhaps the trail of their evolution has been partly obscured by extensive sequence changes in the related domains of plants, fungi and protozoa. As we shall discuss, the fact that three-dimensional structures are more persistent than sequences in an evolutionary sense may allow this puzzle to be solved.

Besides all the evidence that most exons are not evolutionarily mobile, some mobile domains are also clearly not the products of single exons. One large domain, first observed in the fibrinogen molecule of vertebrate animals, is composed of 250 amino acids. In some proteins, the gene for that domain contains multiple introns. Yet none of these individual exons has ever been found without all the others. None of the exons, it seems, has ever shuffled itself out of the domain. Thus, the mere presence of introns in a gene is apparently not enough to cause exons to be mobile. That the vast majority of identified exons has never been found in more than one setting argues against simple and indiscriminate mobility.

There are other examples of movable units that contain introns within their coding sequences. One of the first movable modules to be identified is referred to as a kringle (because it resembles a type of Danish pastry by that name). It is made of about 80 amino acid units and contains three characteristic disulfide bonds. It is quite similar to the Fn2 domain, differing only in the number of residues between its cysteine residues, and some workers do not distinguish between the two. In some of its settings, the gene for the kringle is split by an intron, but so far no one has found half a kringle in any protein.

A further point in favor of the "intron late" theory is that the introns that interrupt coding regions are much more common in plants and animals than among the earliest diverging eukaryotes. No introns at all have been found in primitive eukaryotes, such as *Giardia lamblia*. Furthermore, modular proteins have been identified in plants for which no recognizable counterpart is known in animals, and vice versa. Finally, there is indirect evidence that the modular assembly of some bacterial proteins occurred so recently that they must have evolved without the aid of introns. All this evidence suggests that protein-interrupting introns appeared after the evolution of eukaryotes.

So some exons encode domains, but most exons do not; those that do can often be genetically duplicated and shuffled. The issue of cause and effect between these phenomena raises a thorny problem. Perhaps the evolution of these introns did facilitate exon shuffling. On the other hand, it is not impossible that these introns often delineate the coding regions for domains because that placement is advantageous for intron propagation. If an intron interrupted a sequence coding for a domain, it might survive in that situation (if it did not violate the phase rule noted earlier). It would not spread further, however, because the exons on its boundaries could not stand alone and thus could not be moved independently. Conversely, if the intron lands between regions coding for independently folding units, it can spread to other locations along with the exons. So exon shuffling may be only incidental to the survival of introns.

One way to learn more about the evolution of movable modules is to look at the structure and dispersal of one in particular. Our favorite is Fn3, the fibronectin type III domain. Like kringles, Fn3 units are sometimes split by single introns, but they have never been observed with fewer than their full complement of 90 to 100 amino acids. The two of us independently followed the discovery of Fn3 in various proteins for several years. In early reports the sequence was turning up only in animal proteins, so we were both surprised when, in 1990, workers at Niigata University in Japan reported the existence of an Fn3 domain in a bacterial protein. Our common interest came to light at a meeting in Italy in 1991, whereupon we decided to join forces in a comprehensive inventory of Fn3 occurrences.

To that end, we screened a data base of protein sequences by various means, including a pattern-searching algorithm that one of us (Bork) had devised with Christian Grunwald when both were at the Central Institute for Molecular Biology in East Berlin. We uncovered well over 300 unique occurrences of the Fn3 sequence motif, which established that it was indeed a true domain, should there have been any doubt. The 300 occurrences actually represented 67 different proteins, not counting the same proteins from different species. Sixty of these were from animals and seven from bacteria. None of the sequences identified were from plants, fungi or unicellular eukaryotes.

The obvious questions to be asked were: Did bacteria and animals both inherit the domain from some mutual ancestor, or did one of the groups somehow acquire it from the other? If the domain was present in the common ancestor of prokaryotes and eukaryotes, then why has it not been found in fungi and plants? With the aid of a computer, we aligned all the Fn3 sequences that we could find and constructed a crude phylogenetic tree based on their similarities. Because trying to accomplish that with all 300 sequences was cumbersome, we began by using representative sets of sequences from all the bacterial proteins and from only the most different of the animal ones.

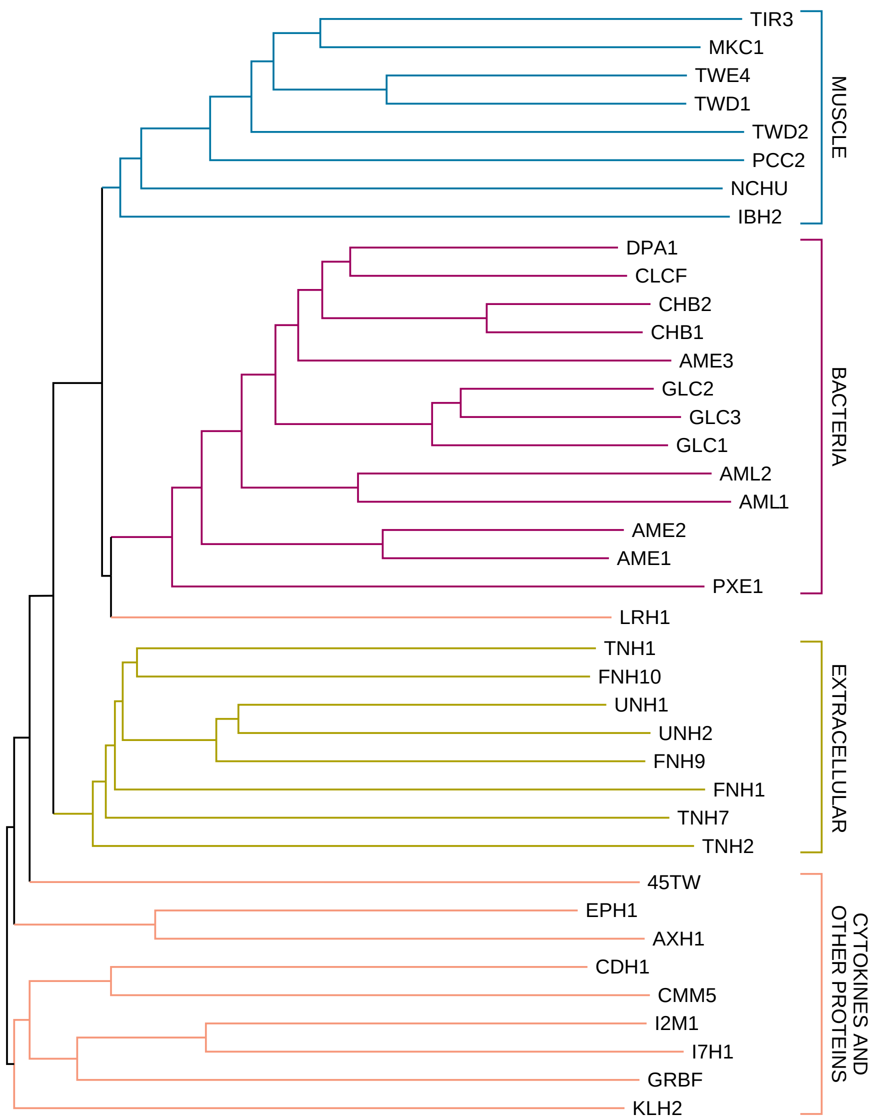
It was soon clear that something was amiss. The bacterial sequences were simply too similar to the animal ones to have descended from a common ancestor two billion years ago. Instead the evidence—including the computer-generated phylogenetic trees—favored the view that somehow the bacteria had acquired the Fn3 domain from an animal source.

There were several reasons to think so. First, it was often the case that an enzyme from one bacterial species carried the domain but that the same enzyme from other species did not, which implied that the domain was structurally and functionally expendable. It must therefore have been a late addition to specific bacteria. Moreover, the Fn3 sequences appeared sporadically and in clusters but always in a characteristic set of extracellular enzymes; if bacteria were losing copies of Fn3 over time, one might expect to find surviving copies in a more varied set of proteins.

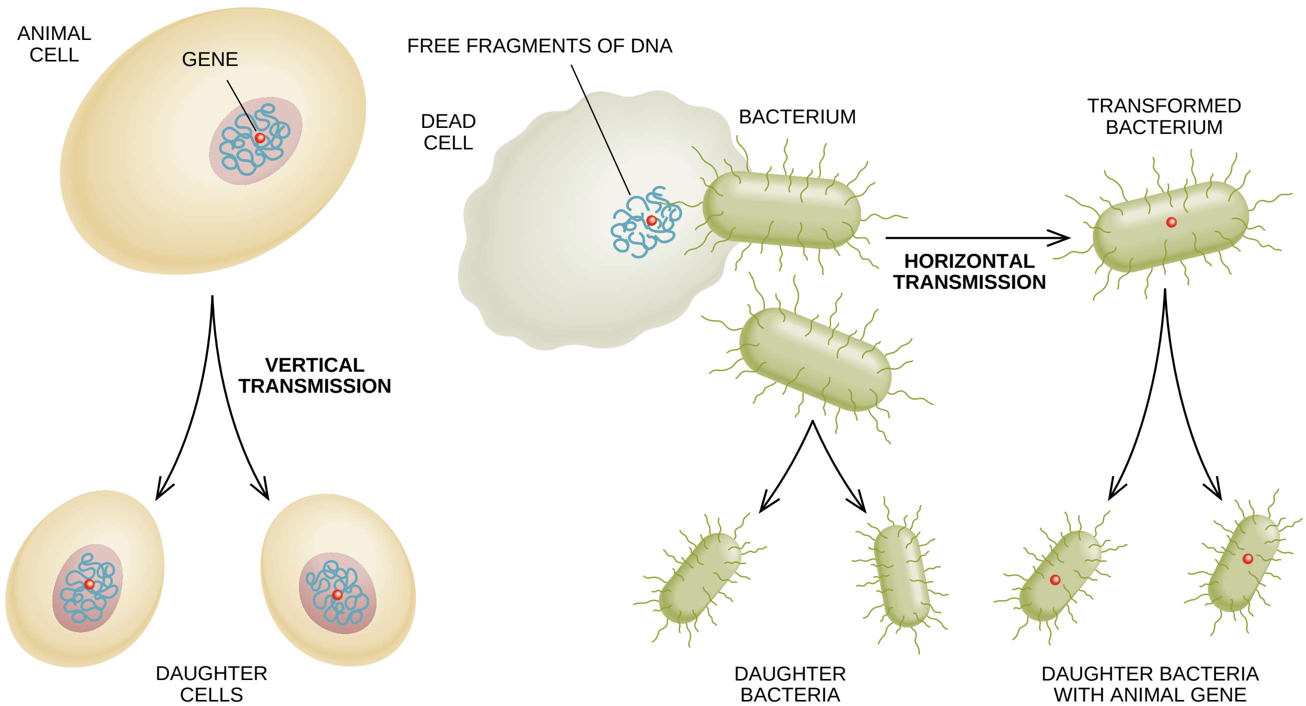
Lastly, although the bacteria that had Fn3 domains were of diverse types, they did have certain features in common. All are soil bacteria that obtain their food from commonly found polymers, such as cellulose and chitin, released during decomposition. Vast numbers of other types of bacteria have been examined, but none of them have the Fn3 domain in any of their proteins. More than half the genome sequence of the intestinal bacterium *Escherichia coli* is known, and not one hint of an Fn3 sequence has surfaced. The same can be said for the large number of studied fungal and plant sequences. If the Fn3 domain had been present in a common ancestor of prokaryotes and eukaryotes, we would have expected its radiation to have continued along the major lines of descent and to be represented in all these groups.

The idea that the gene for a domain can migrate between distantly related organisms may seem outlandish at first. Common experience shows that genes are transmitted vertically, from one generation of organisms to the next. It is nonetheless sometimes possible for genes to be transmitted horizontally as well, not only between species but across distant lineages. Some viruses can pick up small genes from one host and carry them to another host; in rare instances, the transposed gene may assimilate itself into the new host's DNA. Bacteria can be transformed by picking up DNA from their surroundings, which may include decaying animal cells. Many bacteria also carry genes on small rings of DNA, or plasmids, that they can exchange with other bacteria. In theory, all these mechanisms represent opportunities for the horizontal transmission of genes.

Allowing that some bacterium did obtain the gene for an Fn3 domain from an animal source, when did it happen? All that the phylogenetic tree can indicate is that it happened within the past billion years, after the divergence of animals from plants and fungi. To pin down the date, we need to know the average rates of sequence change along both the bacterial and animal lines of




**EVOLUTIONARY RELATIONS of 39 Fn3 domains from various sources appear in this phylogenetic tree. Using a computer, the authors created the tree by comparing the sequences of the domains from animals and bacteria. Surprisingly, the bacterial domains are closely related to some from animals, which suggests that bacteria obtained the genes for those domains from animals.**




**HORIZONTAL TRANSMISSION OF GENES could explain how bacteria obtained domains from animals. Genes pass vertically between generations of cells. Bacteria can also be transformed by the absorption of DNA from their environment. If a bacterium picked up the DNA for a domain from a dead cell, it might have transmitted the gene to its progeny.**

divergence. For the animal proteins, we can estimate the rate by comparing sequences from various creatures whose divergence times are shown in the fossil record. Unfortunately, we do not have comparable information about the bacterial sequences. (Whereas some microfossils corresponding to bacteria have been reported, there is certainly not yet an interpretable evolutionary hierarchy for them as there is for animals.) We can see, however, in both the animal and bacterial proteins a tendency for tandem duplication of Fn3 sequences—that is, within any one protein containing more than one Fn3 domain, the sequences of those domains are often adjacent and usually very similar. That observation implies that the duplication of the DNA for the Fn3 domain must be relatively recent.

The timing of the horizontal transmission and of the genetic duplications is critical to our understanding of how these genetic units spread. So far as is known, contemporary bacteria do not have introns in their protein-coding gene sequences. If they ever did have introns in their coding sequences, how long has it been since those introns disappeared? Unless it was fairly recently, the genes for Fn3 must have been spreading without the aid of introns.

One likely possibility is that the Fn3 domain is being spread among soil bacteria by a promiscuous bacteriophage (bacterial virus) or plasmid. We hope eventually to catch the act of transmission in progress by finding a phage that carries a gene for an Fn3 domain. Now that a number of bacterial Fn3 sequences have been found, it may be possible for us to synthesize short tags of DNA that will bind to the Fn3 genetic units. Used in conjunction with the DNA amplification technique called the polymerase chain reaction, those tags could help identify the genes in bacteriophages or other vectors.

Of course, where the Fn3 domain originated is still a mystery. Did it first appear in animals? Or are we just unable to identify its ancestral forms by sequence comparison? Three-dimensional determinations show that the Fn3 structure is suspiciously similar to immunoglobulin domains. Three-dimensional analyses by x-ray crystallography and NMR have made it possible to trace the immunoglobulin domain back to proteins in prokaryotes, including PapD, a “chaperone” that helps other proteins fold, and also to a bacterial enzyme that digests cellulose. It is interesting that the immunoglobulin domain, as originally defined, contains a disulfide bond that pins its two sides together, but the more primitive forms, some of which persist in vertebrate animals today, lack that bond. It is these primitive forms that are most similar to the Fn3 domain.

We can anticipate that other instances of pirated modules will be uncovered. By our census, the Fn3 domain occurs in about one out of every 50 animal proteins (50 out of 2,500 known animal sequences, independent of species redundancies). We estimate that about 25 modules are very commonly strewn throughout animal proteins, much as Fn3 is. More than 100 others are found in more than one setting but less frequently than the first group. Tracing the pedigree and diaspora of these modular units is a major challenge that should reveal much about every aspect of the evolution of all living things.

FURTHER READING

THE GENEALOGY OF SOME RECENTLY EVOLVED VERTEBRATE PROTEINS. Russell F. Doolittle in *Trends in Biochemical Sciences,* Vol. 10, No. 6, pages 233–237; June 1985.

PROTEIN MODULES. Martin Baron, David G. Norman and Iain D. Campbell in *Trends in Biochemical Sciences,* Vol. 16, No. 1, pages 13–17; January 1991.

MODULAR EXCHANGE PRINCIPLE IN PROTEINS. Lázló Patthy in *Current Opinion in Structural Biology,* Vol. 1, No. 3, pages 351–361; June 1991.

MOBILE MODULES AND MOTIFS. Peer Bork in *Current Opinion in Structural Biology,* Vol. 2, No. 3, pages 413–421; June 1992.

PROPOSED ACQUISITION OF AN ANIMAL PROTEIN DOMAIN BY BACTERIA. Peer Bork and Russell F. Doolittle in *Proceedings of the National Academy of Sciences,* Vol. 89, No. 19, pages 8990–8994; October 1, 1992.

# Electrorheological Fluids

*Some liquids solidify instantly when exposed to an electric field. Such protean materials may give engineers quicker, more adaptive machines*

by Thomas C. Halsey and James E. Martin

The T-1000—a nearly indestructible robot that was brought to life in the movie *Terminator 2: Judgment Day*—could deftly transform from liquid to solid. When pierced by a bullet, the robot's liquid-metal skin flowed into the hole; when broken into small pieces, the T-1000 melted and then congealed into its original form. Electrorheological fluids promise some of the adaptability that the film's producers imagined.

An electrorheological fluid is a substance whose form changes in the presence of electric fields. Depending on the strength of the field to which it is subjected, an electrorheological fluid can run freely like water, ooze like honey or solidify like gelatin. Indeed, the substance can switch from one state to another within a few milliseconds. Electrorheological fluids are easy to make; they consist of microscopic particles suspended in an insulating liquid. Yet they are not ready for most commercial applications. They tend to suffer from a number of problems, including structural weakness as solids, abrasiveness as liquids and chemical breakdown, especially at high temperatures.

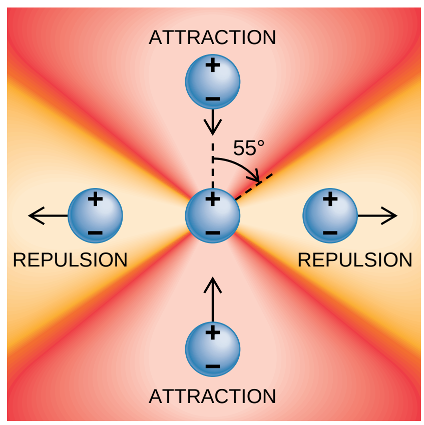
Researchers have recently learned to make fluids with less abrasive ingredients and more stable chemistry. Structural weakness, however, remains a serious impediment. If practical electrorheological fluids are to be designed, it is important to have a firm understanding of the electrical interactions among suspended particles that give rise to strong solids.

Electrorheological fluids have fascinated scientists not only because of their useful physical properties but also because of their complexity. The particles in an electrorheological fluid organize themselves into a variety of intricate structures. At one extreme, when the fluid flows freely, the particles move independently of one another. At the other extreme, when the fluid solidifies, the particles link together to form thin chains and thick columns that are visible to the naked eye.

In studying electrorheological fluids, we are thus probing the behavior of matter somewhere between the typical liquid and solid, quite distant from ordinary liquid-state or solid-state physics. Recent research has provided many new insights into the interaction between particles in these unusual states. It remains to be seen whether such findings will contribute to the design of stronger fluids.

Investigators began taking an interest in the electrorheological effect around 1947, when Willis M. Winslow obtained the first patent on an electrorheological fluid after some eight years of observations. It was quickly discovered that something as simple as cornstarch dispersed in corn oil could make a good electrorheological fluid. This edible home recipe exhibits the electrorheological effect when it is subjected to a strong electric field in the range of 1,000 volts per millimeter. The effect can be produced, for example, by placing the fluid between two metal plates one millimeter apart and charging the electrodes to 1,000 volts. That voltage may seem high, but the power requirement is actually small. Because the fluid is an insulator, no appreciable electric current flows through it. Consequently, the charge on the plates rapidly builds up to a high voltage, just as thousands of volts of static electricity can easily accumulate on one's skin.

Winslow's patented fluid was a very exciting discovery, and the future of electrorheological fluids looked bright. Automotive engineers could imagine, for instance, constructing an electrorheological clutch. The fluid inside this clutch would couple a car's engine to its drive shaft, which delivers power to the wheels. When solidified, the fluid




**INTERACTION between charged particles is the key to the electrorheological effect. A particle that has a positively charged pole and a negatively charged pole can either repel or attract another particle polarized in the same manner (*above*). When particles are side by side, the repulsion between like poles forces them apart. When particles align vertically, the attraction between unlike poles brings them together. In the illustration at the right, a polarized particle in the repulsive zone (*orange*) of another particle circles around into the attractive zone (*purple*). If many polarized particles suspended in a fluid interact in this way, the particles will form chains, and the fluid will exhibit the electrorheological effect; that is, the fluid will thicken or even solidify.**

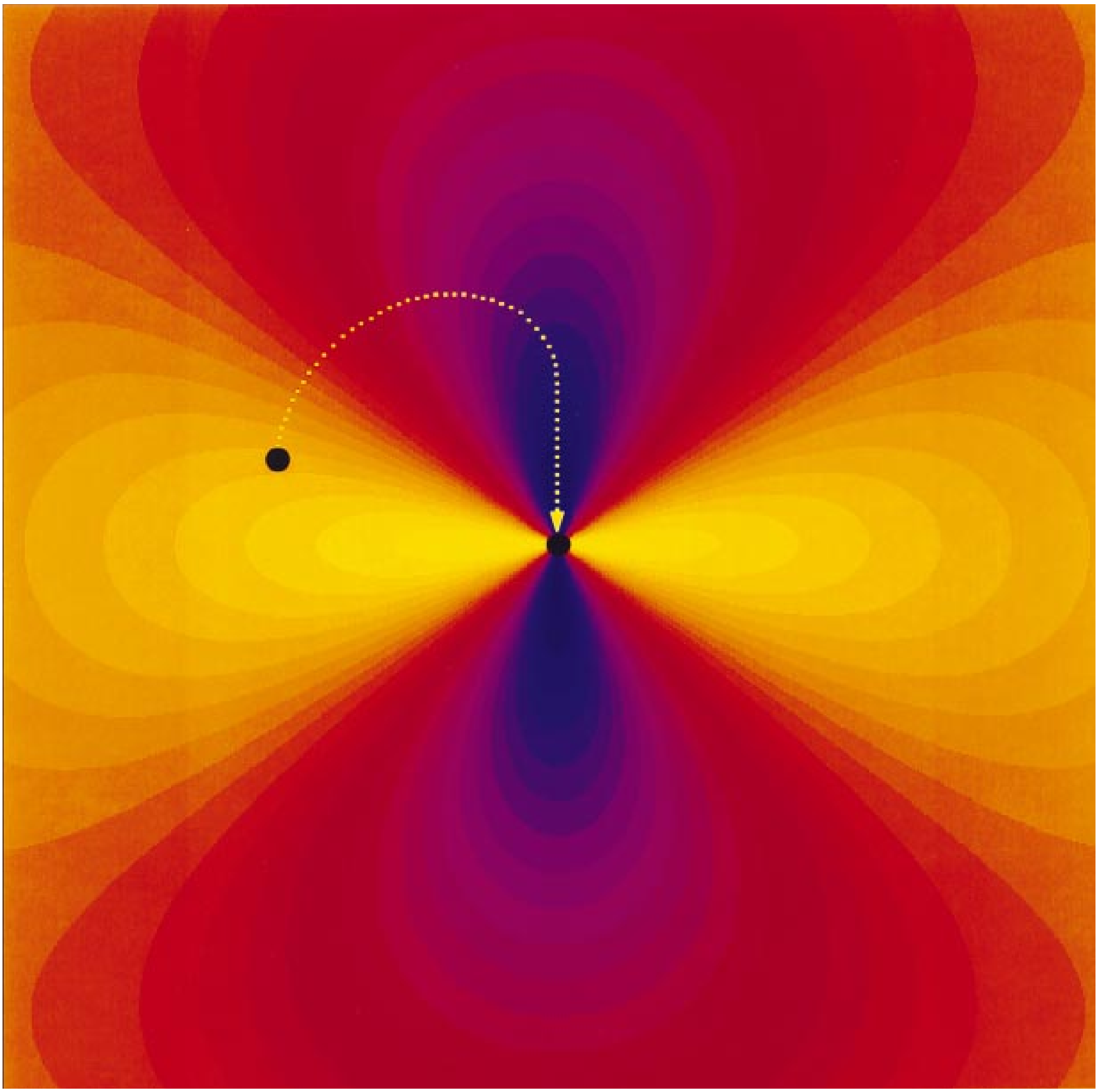
THOMAS C. HALSEY and JAMES E. MARTIN have made major contributions to the field of electrorheology during the past few years. Since 1987 Halsey has been an assistant professor of physics in the James Franck Institute at the University of Chicago. After earning his Ph.D. in physics from Harvard University, he became a Sloan Research Fellow and a Presidential Young Investigator. In 1986 he served as a visiting scientist in theoretical physics at the Center for Nuclear Studies in Saclay, France. Martin is a senior member of the technical staff at Sandia National Laboratories. He joined the staff at Sandia in 1981 after earning a Ph.D. in physical chemistry from the University of Washington.

would force the drive shaft to rotate. As it liquefied, the fluid would allow the engine to disengage from the drive shaft and spin freely, as if in neutral gear. Such a clutch would have few parts to wear or fail. And with a response time of a few milliseconds, it would make purely mechanical clutches look slow. It was also hoped that electrorheological fluids would lead to valveless hydraulic systems, in which solidifying fluid would shut off flow through a thin section of pipe.

Those ideas are still attractive, but the truly exciting possibilities for electrorheological fluids involve adaptable devices that take advantage of the fluid's continuously variable viscosity. Because the flow rate of an electrorheological fluid can be controlled by the application of an electric field, novel concepts such as the adaptive shock absorber have drawn interest from the automotive industry as well as from proponents of magnetically levitated trains. (Although such trains do not ride on rails, adaptive shock absorbers could damp the effects of quick stops and starts.)

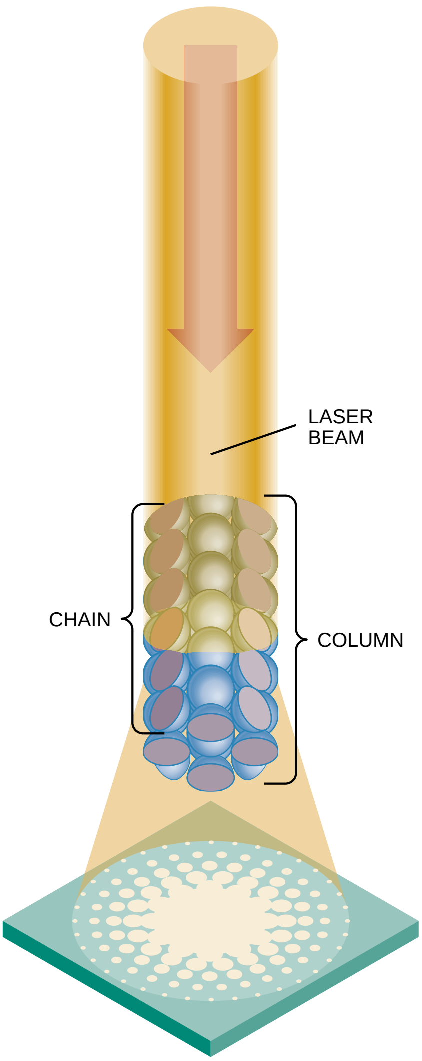
An ordinary shock absorber consists of a cylinder with a piston inside that when compressed pumps a thick, viscous oil through a small orifice. The fluid absorbs the force of an impact by its slow, sticky motion, and a larger orifice then enables it to return more quickly into the shock absorber during the rebound. Unfortunately, the viscosity of oil is very sensitive to temperature. On a cold day the oil becomes thicker, so the shock absorber is a bit stiff. Furthermore, because the oil heats and becomes thinner with repeated compression, the shock's damping ability nearly vanishes on a long, bumpy road, where it is needed most. One way to compensate for these difficulties is to introduce a mechanical system that adjusts the size of the shock absorber's orifice as the need arises.

Electrorheological fluids offer the possibility of a shock absorber that provides response times of milliseconds and does not require mechanical ad-



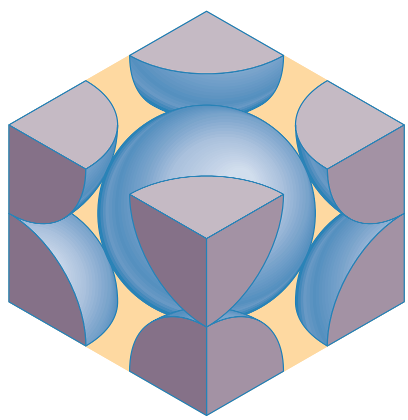
justments. A shock absorber can be made by filling a piston with electrorheological fluid and adding electrodes and a microprocessor. As the piston moves, the microprocessor rapidly modulates the voltage across the electrodes to change the fluid's viscosity. Such fast control would make it possible to increase the fluid's thickness in midstroke, so as to damp the effect of a pothole. The fluid could then be made thinner again for a quick recovery. A single shock absorber could thus adapt to a variety of vehicles and operating conditions.

Yet existing electrorheological fluids fail when subjected to heavy punishment. The compounds are often too soft as solids, and the high voltages required to solidify them, if frequently applied, can heat volatile liquid ingredients and eventually boil them away. To address these problems, we are investigating how the electric field causes solidification. Ideally, we would like to have fluids that become strong solids at lower voltages than are now needed.




Investigators have known for some time that the origin of the electrorheological effect lies in a phenomenon termed electrical polarizability. In an electric field—between, say, a pair of charged metal plates—the positively charged protons inside a particle will be attracted toward one electrode while the negatively charged electrons will be attracted toward the opposite electrode. The result is that the positive and negative charges shift somewhat inside the particle, each in an opposite direction. A particle in an electric field thus becomes an electric dipole, whose positive and negative poles are analogous to a magnet's north and south poles. The strength of this effect varies according to the electrical polarizability of the substance. If the polarizability is large, then large amounts of electric charge shift over large distances toward the particles' opposite poles. If the polarizability is small, then the particles hardly respond at all to an applied electric field.

When the fluid contains many polarized particles, the electrical interactions become complicated. The simplest situation involves two particles in a strong electric field, each particle having its positive pole on top, for argument's sake, and its negative pole on the bottom. The force of electrical attraction or repulsion between these two particles depends on the way they approach each other. If the first moves up directly beneath the second, then the positively charged top of the first is closest to the negatively charged bottom of the second, and, since opposites attract, the two particles feel a net attractive force.



**LASER LIGHT shining through a solidified electrorheological fluid produces a pattern that offers clues about how the particles of the fluid are organized. In this experiment each particle acts like a lens because each is made of glass and is much larger than the wavelength of light. A pattern of circles implies a three-dimensional structure known as the body-centered tetragonal lattice (*above*).**

If the first particle somehow manages to come alongside the second, aligning like the letters in the Roman numeral II, then the negatively charged bottom of the first is closest to the negatively charged bottom of the second, and likewise for the positively charged tops. Because likes repel, the two particles feel a net repulsive force. Of course, if the particles approach each other at some other angle, the attractive or repulsive force is less than it is in the extreme cases. In general, if a line connecting the two particles is within 55 degrees of the vertical, the particles attract; otherwise, they repel.

Even if initially repelled, the mobile particles tend to circle around one another. Their trajectories curve, like those of comets, in response to the forces attracting or repelling them. At the same time, no matter in which direction the particles are headed, their poles point either up or down toward the electrodes. Together these forces bring virtually all particles into a configuration in which oppositely charged poles are close together [*see illustration on preceding page*]. Then the attractive forces take over, and the particles stick together. Successive particles line up end to end in much the same way to create long chains, like pearls on a string. This behavior is analogous to the familiar alignment of iron filings along the field lines of a bar magnet. The chains of particles in an electrorheological fluid quickly grow from one end of the fluid container to the other. Formation of such chains is responsible for the solidification of the fluid.

Some particles readily form chains when suspended in one kind of fluid but not when suspended in another, and researchers are still working to sort out the messy details. One potential complication is that the particles are immersed in liquid whose molecules can also be polarized. If the polarizability of the liquid is exactly the same as that of the particles, then the polarization of the liquid masks that of the particles, and chains will not form. In general, electrorheological fluids work best when the particles polarize more easily than do the molecules of the fluid.

Still, chains are only an intermediate step in the evolution of structure in electrorheological fluids. Single chains slowly migrate toward nearby parallel chains to combine as thicker columns, which combine to form still thicker columns, and so on. What is the ultimate solid structure? Recent theoretical work and experimental observations have revealed the answer.

First, an analysis of the complex electric fields around large numbers of particles offers clues about the structure we might expect. As we have already seen, dipolar particles repel one another when side by side, so the appearance of columns made of parallel chains at first seems mysterious. Why don't such columns disintegrate into their constituent chains? This does not happen because a particle is attracted to or repelled from a neighboring chain not only by the force from the few nearby chained particles but also by the force from many distant particles in the chain. Columns stay together because the dipolar force is long-range, and particles respond to the sum of forces from all other dipoles in the fluid.

To understand why the columns hold together, consider a "test" particle placed near a chain. The test particle will be repelled by the nearest particle in the chain, just as side-by-side particles repel one another. But the test particle will be attracted by the chained particles well above and below it because the test particle will be in the 55-degree attractive zone of most of the particles from the top and bottom of the chain. To be even more precise, the test particle is drawn to a position slightly above one particle in the chain and slightly below another. Consequently, if many test particles are placed near the chain, they will be attracted to it and each other and can form a second chain and a third, and so on. In this way, the dipolar forces between particles encourage the chains to form columns in which the particles in adjacent chains are staggered.

Tian-jie Chen, Robert N. Zitter and Rongjia Tao of Southern Illinois University recently confirmed the existence of such a structure. In their experiment, they used an electrorheological fluid made of glass spheres with a relatively large diameter of 40 microns, suspended in a silicone oil. Immediately after the electric field was applied, chains formed. Within a few minutes the chains aggregated into columns with a roughly circular cross section approximately six millimeters in diameter.

To determine the structure within these columns, the investigators utilized the focusing properties of the glass spheres, treating the columns as stacks of lenses that could transmit light. Light propagated from sphere to sphere until it exited from the column. The last layer of spheres acted as a set of light sources and created a pattern of overlapping spots [*see illustration on page 60*]. By analyzing this pattern, the Southern Illinois group was able to infer that the structure of their solidified electrorheological fluid was a common, orderly arrangement of spheres (a three-dimensional lattice) known as the body-centered tetragonal structure, in which the particles in adjacent chains are indeed staggered.

Although the attractive forces between chains are strong enough to create such a lattice, they are too weak to account for the relatively rapid aggregation (over a few minutes) observed by the Southern Illinois group. Reality, it seems, is more complicated than any neat analysis of electrical forces.

In experiments, columns form faster than expected because the chains have imperfections. For at least two reasons, particle chains are not perfectly straight, ordered structures. First, chains sometimes fail to span the entire distance between the electrodes, and, second, they are shaped, in part, by the Brownian motion of the particles. Brownian motion, first observed in 1827 by the Scottish botanist Robert Brown, is a ceaseless random motion of particles suspended in a fluid; it is caused by collisions between a particle and hordes of fluid molecules. In an electrorheological fluid the suspended particles dance chaotically about their average position in the chain as they are struck by fluid molecules hitting them from every direction. Thus, although the chain may be straight on average, at any given time it is kinked by Brownian motion. This slight deformation increases the force of attraction between chains and catalyzes the aggregation of chains into columns.

To obtain a better understanding of column growth, the authors, along with Judy Odinek of Sandia National Laboratories, employed an optical method quite different from the technique developed by Chen, Zitter and Tao. Instead of directing light through chains made from relatively large glass spheres, we aimed a laser beam in a direction perpendicular to the chains and used much smaller spheres—less than a nanometer in diameter. Light scattered from columns made of such particles carries with it information concerning the spacing and width of the columns. In general, the exiting light makes a pattern like a figure eight. At first, the figure-eight pattern is large and diffuse. As the columns thicken and the distance between them increases, the light traveling through the fluid encounters fewer obstacles, so scattering decreases and the lobes in the figure eight become smaller and brighter.

This rough rule lets us interpret the patterns of light seen in our experiments. When the electric field is off, and there are no chains or columns in the sample, the light is scattered uniformly. But when the field is turned on, the light exiting the sample begins to concentrate into two lobes. As time progresses, the light becomes brighter near the center of the figure eight. This brightening indicates less scattering as the columns thicken and more space opens up between them. Column growth is initially very fast, but after a few minutes its progress virtually stops, because the attractive force between the columns is too weak for continued growth. Because calculations show that the solidifying fluid may get stronger as the columns

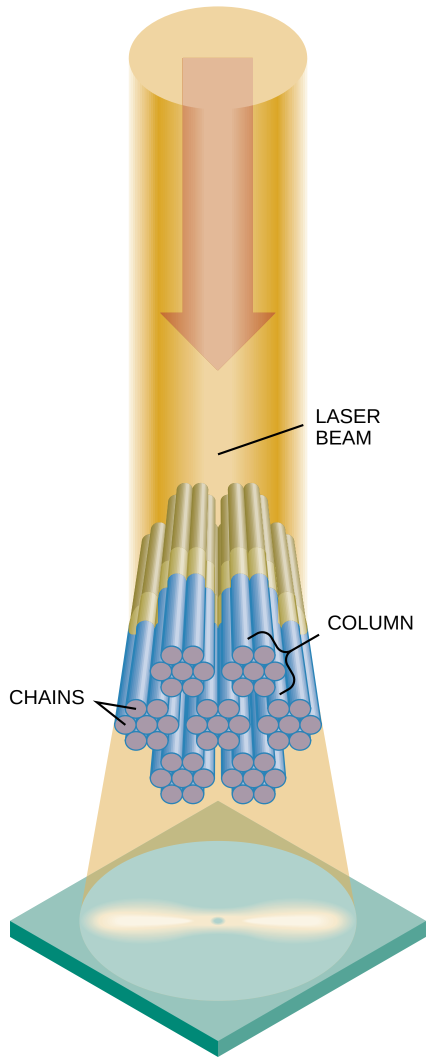
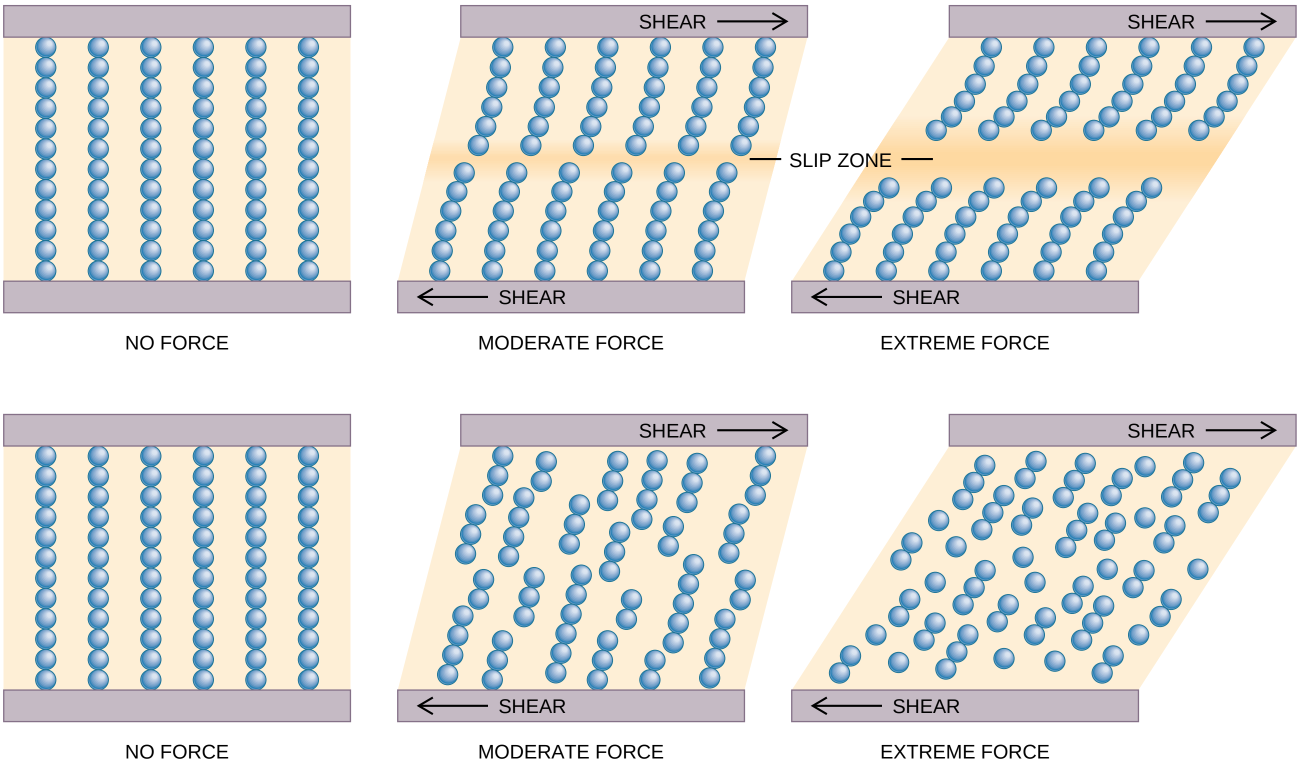



**FIGURE-EIGHT PATTERN appears when light penetrates an electrorheological fluid consisting of particles approximately as small as the wavelength of light. This pattern indicates that the particles have organized themselves into columns. As the distance between the columns increases, the figure eight becomes smaller.**




**TWO THEORIES describe how chains of particles might inhibit the flow of electrorheological fluids. In the Klingenberg and Zukoski theory (*top row*), chains bend and stretch as their ends adhere to shearing electrodes. The chains finally break in the center of the fluid, and this "slip zone" grows as the shearing force increases. In the theory of Halsey, Martin and Adolf (*bottom row*), chains float throughout the fluid. As the shearing force increases, the chains tilt and fragment.**

thicken, these measurements demonstrate that keeping the electric field on for periods up to several minutes may improve the mechanical properties of electrorheological fluids.

These fluids, after all, need to withstand mechanical forces. Knowing their static structure is not enough; we must also understand how the structure behaves when squeezed or stretched. Such issues fall into the domain of rheology, which is the study of the flow properties of materials. (Almost all materials, even solids, flow under sufficiently high pressure, although sometimes they do so by imperceptible amounts.)

In general, the stiffest materials flow very little, whereas the weakest ones flow at the slightest provocation. One variable affecting the flow rate of a fluid is the stress, or force, applied to it. For simple fluids such as water, the flow rate is proportional to the stress: water flows faster when you pump harder. On the other hand, the flow rate is inversely proportional to the viscosity, or thickness, of the fluid. Viscous fluids such as thick oil flow slowly at a given stress, compared with water, for instance.

But electrorheological fluids in an electric field are not so simple. Their behavior resembles that of a family of materials known as Bingham plastics. Below a certain threshold of stress, a Bingham plastic does not flow at all. Beyond this threshold, the flow rate increases in proportion to the applied stress minus the threshold stress. Most of us encounter a Bingham plastic every morning, in the form of toothpaste. This type of rheology creates the "plug" flow that enables us to squeeze a neat cylindrical plug of nonrunning paste onto a toothbrush. Like toothpaste, electrorheological fluids resist a certain amount of stress but begin to flow when more force is applied.

Stress often builds up in an electrorheological fluid when the electrodes start to move—one plate toward the right, say, and the other toward the left. This is called a shearing force. If we apply a small force, the chains of particles in the fluid will tilt and stretch but not break. Yet when the chains are tilted, the attractive force between particles is diminished, because the chains are no longer aligned straight up and down in the electric field. Instead the chained particles move toward a slanted configuration that gets ever closer to forcing the particles side by side. If the chains tilt too far, the particles slip out of the zone in which they mutually attract, the chains break and the material flows. The amount of stress at failure is called the yield stress.

For applications such as the electrorheological clutch, maximizing the yield stress is very important. A fluid with a high yield stress will withstand a high torque, or rotational force, before flow begins and the clutch slips. One way to increase the fluid's resistance to stress is to apply a higher voltage. The amount of separation achieved between positive and negative charges on the particle dipoles increases in proportion to the voltage applied, and the attractive force between particles goes up as the square of this charge separation. The yield stress is therefore proportional to the square of the applied voltage. Increasing the voltage, however, also increases power consumption and heats the electrorheological fluid. If the fluid gets too hot, it might decompose. Hence, in designing strong electrorheological fluids, one goal is to minimize the voltage needed to achieve a given yield stress.

The voltage needed depends largely on particle polarizability. A few years ago it was thought that the strength of a solidified fluid would not increase dramatically with polarizability. We know better now, thanks to studies of highly polarizable ceramic particles conducted by Douglas B. Adolf, Robert A. Anderson, Terry J. Garino and Bradley G. Hance of Sandia National Laboratories. The investigators have shown that strength grows dramatically as polarization increases. There is much optimism that highly polarizable particles will lead to very strong electrorheological fluids.

Electrical conductivity is also a concern. Differences in the conductivity of the particle and the liquid solvent are especially significant if the applied electric field is on for more than a hundredth of a second or so. Consider nonconducting particles suspended in a weakly conducting liquid. Immediately after the field is turned on, the particles will polarize. But charges will also begin to flow through the liquid and will accumulate on the surfaces of the polarized particles. Shortly thereafter, the interactions between particles will be determined not so much by their intrinsic polarization as by clouds of charge encircling each particle in the liquid. The basic force between particles is still dipolar, but the microscopic origin of this force is different, and the yield stress may be different.

Other problems arise with conducting particles in an insulating liquid. When conducting particles polarize, negatively charged electrons do not just shift inside molecules but actually migrate from molecule to molecule across the particle, leading to a strong polarization and strong attractive forces. But if the particles contact each other, charge flows between them. In that case, one of the particles will have a net positive charge, and the other will be negatively charged. Instead of linking together, these particles will migrate toward opposite electrodes, since the electrodes exert a more powerful pull than that of any single small particle. This phenomenon has been seen in suspensions of conducting particles and is colloquially referred to as the "bouncing ball" effect. The exchange of charges between conducting particles can be prevented by coating the particles with a thin insulating layer. Such coatings allow the particles to become highly polarized, and the result is that the fluid displays a strong electrorheological effect.

A high yield stress is not the only useful property of an electrorheological fluid. In some applications, the fluid should yield at a low stress, like toothpaste, yet continue to resist motion as the stress increases. To control such a fluid's varying thickness, we need to learn more about the behavior of electrorheological fluids after they yield and begin to flow. We and several other research teams are therefore investigating what kinds of structures evolve as fluids are stressed. We are examining how these structures inhibit the flow of the liquid in which they are suspended and how this flow then distorts and weakens the solid structures so that the liquid keeps flowing.

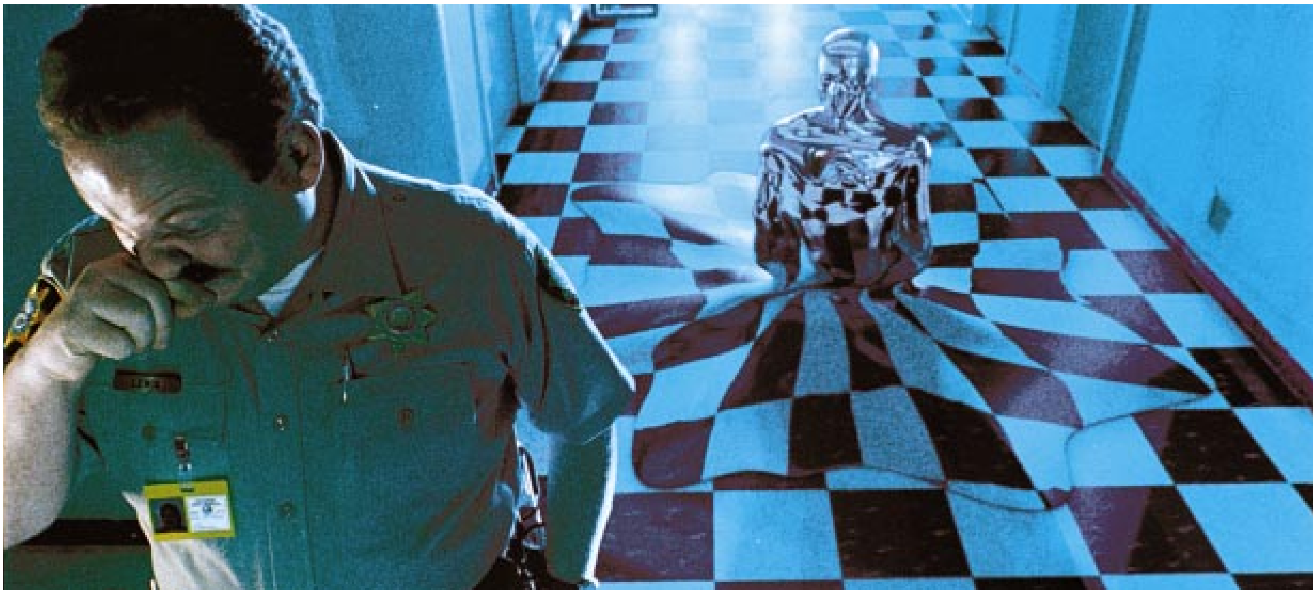
So far there is no comprehensive theory. One group has proposed a model of electrorheological flow based on observations of a cornstarch fluid, whereas our group, experimenting with other fluids, has a different explanation for their thick, resistant flow.

Daniel J. Klingenberg, now at the University of Wisconsin, and Charles F. Zukoski of the University of Illinois studied the flow of a cornstarch electrorheological fluid at stresses larger than the yield stress. They found that shearing the electrodes gave rise to two kinds of structure. Near the plates, tilted columns of particles formed. In the central region between plates, a "slip zone" developed in which the chains were broken and the fluid moved freely. As the rate of shearing increased, the slip zone increased in width and lubricated the relative motion of the tilted columns adhering to the plates.

Working with Adolf, we have proposed an alternative picture of shearing in electrorheological fluids in which the adhesion of columns to the electrodes is not dominant. If the columns are free to drift, the attractive forces between particles tend to align the columns perpendicular to the electrodes. Yet the flowing fluid tries to rotate them so that they are parallel to the electrodes and the direction of flow. These forces arising from the fluid flow are called viscous forces. At some angle of column tilting, the viscous forces that are trying to turn chains parallel to the flow will balance the electrical forces that are attempting to keep them perpendicular. We expect to see such tilted columns throughout the fluid as it yields to shearing stress and begins to flow.

We also expect the columns or chains to shorten as the rate of shearing increases. Because a long chain has more surface area than a short one, the viscous forces act more strongly on a long chain. In fact, the strong viscous forces will rotate very long chains beyond the angle at which the particles attract one another. Therefore, the chains will break up as the particles begin to repel. This phenomenon sets an upper limit on the length of the chains that can appear at a given shearing rate in a given electric field. These relatively short chains will nevertheless dramatically increase the viscosity of the fluid. The situation is analogous to pumping water through a pipe filled with pine needles. The fact that the water must

**SOLIDIFICATION occurs in stages when an electric field is applied to a disordered fluid (*top*). Particles take only milliseconds to form long "pearl" chains (*middle*). A few seconds later the chains can be seen aggregating into columns (*bottom*). These photographs show silica particles one nanometer in diameter.**



**ULTIMATE IN ADAPTIVE MATERIALS was portrayed in *Terminator 2.* The Terminator camouflages itself as a tile floor, changing readily from solid to liquid. The technology of electrorheological fluids is, of course, far less sophisticated.**

push its way through a fine mesh increases the force necessary to achieve a certain flow rate, so water combined with pine needles has a high viscosity, just as electrorheological fluids do beyond the yield stress.

The behavior of many electrorheological fluids can probably be explained by a combination of the model of Klingenberg and our theory. Near the electrodes, a relatively quiescent region similar to the rest state of the solidified electrorheological lattice will form. In the intermediate region, this structure will break up into shorter, tilted chains of particles.

Electrorheological fluids today are no longer the fairly simple mixtures they were in the 1940s. In addition to the suspending liquid and the particles, they often contain a dispersant and an activator. The dispersant keeps the particles from sticking to one another in the absence of an electric field. Without the dispersant, particles sometimes clump together or fail to disperse even when the strength of the electric field is zero. Such temperamental fluids are of little use. The activator, which is often water or sometimes alcohol, contains impurities—often dissolved salts. The mechanism by which the activator works is not entirely clear, but it is generally believed that water, repelled by an oily suspending liquid, collects on the particle surfaces. The dissolved salts then polarize in an electric field, adding their charge to the particles' intrinsic polarization.

These standard formulations have two general problems, aside from structural weakness. First, because the strength of the electrorheological effect increases with particle size, large particles—with diameters much greater than one nanometer—are preferred. But large particles tend to sink to the bottom of the fluid. (Also, fluids containing large particles are more abrasive than are suspensions of smaller particles.) The second problem is that an activator such as water tends to evaporate or undergo chemical reactions with the other ingredients, leading to a short operating life, especially at high temperatures.

The first generation of electrorheological fluids, developed by Winslow, was kept moist by exposure to humid air and contained gritty silica particles. About 10 years ago James E. Stangroom and a group at the Cranfield Institute of Technology in England developed a second generation of less abrasive fluids with particles made of polymers but still containing water. The next step, which is now under way, is to create a third generation of water-free fluids with small, highly polarizable particles.

Meanwhile the machines that might use electrorheological devices have also changed since the 1940s. The automotive industry is still interested in adaptive materials with millisecond response times, but so is another industry that did not exist half a century ago. Today simple robots do much of our industrial work. If the recent revolution in semiconductor technology is to lead to the robotics revolution envisioned by several generations of science-fiction writers, adaptive materials that couple fast microprocessors to equally fast mechanical controls will be crucial. A robot agile enough to perform feats that are fairly routine for human beings—say, catching a baseball—may need a hydraulic system that takes advantage of the rapid response of electrorheological fluids.

We do not need robots to replace human athletes, but imagine, for instance, a computer-controlled device that could extrude the fine metal wire used to connect integrated circuits. When winding this delicate wire on a spool, the robot would have to control tension precisely: the wire might droop or break unless the machine quickly sensed the slightest change in tension and responded. Such a device, with a fast electrorheological clutch, has already been tested in prototype. Others may soon follow—machines, in effect, with superhuman "joints" that instantly lock or loosen with extremely high precision. They would be no match, of course, for the T-1000 Terminator but would be tough competitors in the technological marketplace.

FURTHER READING


INDUCED FIBRATION OF SUSPENSIONS. W. M. Winslow in *Journal of Applied Physics,* Vol. 20, No. 12, pages 1137–1140; December 1949.

ELECTRORHEOLOGICAL FLUIDS AS COLLOIDAL SUSPENSIONS. Alice P. Gast and C. F. Zukoski in *Advances in Colloid and Interface Science,* Vol. 30, Nos. 3–4, pages 153–202; December 1989.

ELECTRORHEOLOGICAL FLUIDS. T. C. Halsey in *Science,* Vol. 258, pages 761–766; October 30, 1992.

# Water-Pollinated Plants

*Once thought to be mere aberrations of nature, these flowering aquatic species provide evidence for the evolutionary convergence toward efficient pollination strategies*

by Paul Alan Cox

On a sunny October day in 1787, the botanist Filippo Cavolini rowed a boat in the Bay of Naples, continuing a pursuit that had so far proved futile. Four years previously, divers had brought him a species of eelgrass, *Zostera marina,* that bore fruit. Cavolini had immediately realized that the plant was special: it must flower, unlike the seaweeds and other marine algae he had collected. But he had yet to discover its blossoms. Cavolini's quest finally ended that day as he reached a *Zostera* population that grew in the bay. He became tremendously excited to find "a spike [flower cluster] not unlike that of a grass floating on the water."

Later, examining the pollen under his laboratory microscope, Cavolini was stunned to see that the grains were not small and round, like those of terrestrial plants: "I found it [the *Zostera* pollen] different from that of other plants, being oblong like little eels, which, with a sudden and brisk motion, exploded and scattered their sperm in a twinkling of an eye." Based on this observation, he drew a startling conclusion: *Zostera* must reproduce by pollination in the open water.

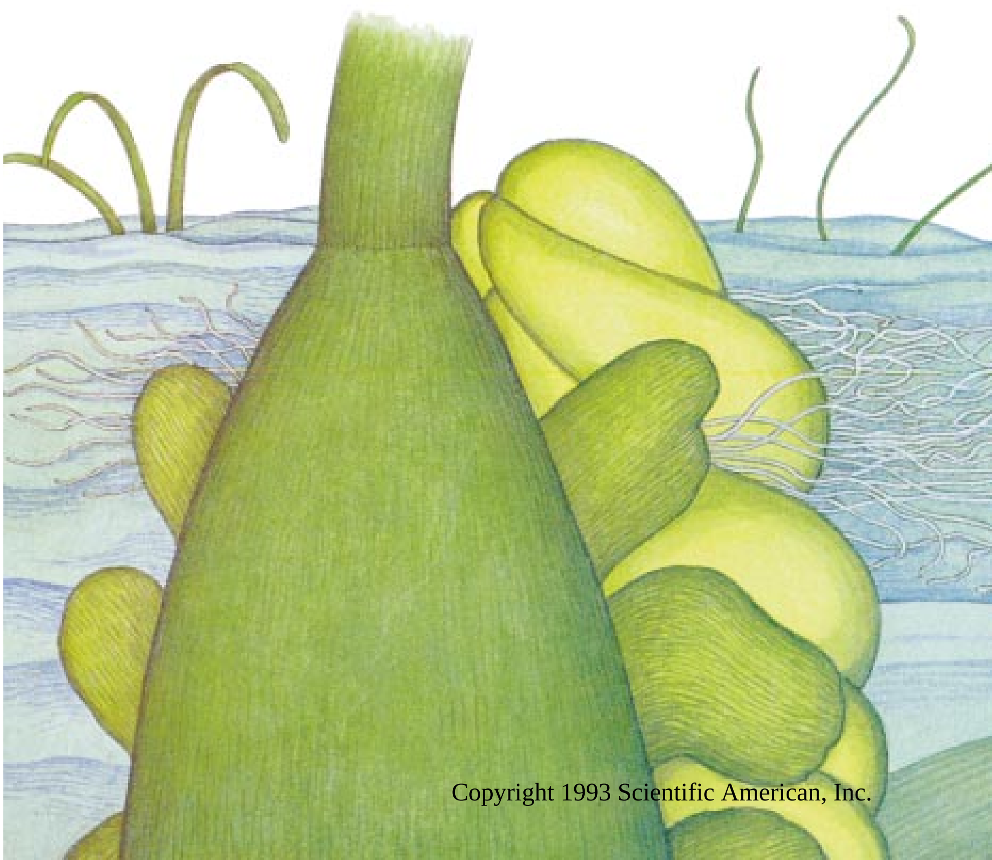
Cavolini's discovery was remarkable. Because water causes the pollen of terrestrial plants to burst, botanists had regarded the aquatic environment as inimical to pollination. Pollen, which transports and protects a plant's genetic information, is a necessary adaptation for flowering plants on land. These plants use wind and animals to disseminate their gametes over dry terrain.

But Cavolini's finding of water as a pollination vector has recently been reconsidered in light of search theory and diffusion physics. The results indicate that hydrophilous, or water-based, pollination systems cannot be considered mere quirks of natural history, as previously thought. Rather water-pollinated plants should now be viewed as compelling cases of convergent evolution, because unrelated hydrophilous species have evolved similar strategies that promote efficient pollination. Although they are uniquely adapted to their watery environment, their ancestors can be ultimately traced to flowering plants that reinvaded the water. Hydrophilous plants thus retain many features of the reproductive systems of their terrestrial counterparts, much as whales and porpoises kept their mammalian traits after their return to the ocean.

Water pollination should not be confused with reproduction in such nonflowering plants as algae, ferns, mosses and liverworts. Although these species require water—say, dewdrops or waterfall spray—they do not produce pollen; they simply disperse their motile gametes into the water. Nor is hydrophilous pollination similar to the aerial pollination of other aquatic flowering plants, such as water lilies, that raise their flowers above the surface to where insects or wind can pollinate them.


PAUL ALAN COX is dean of general and honors education and professor of botany at Brigham Young University, where he received his B.S. After earning his master's at the University of Wales and a Ph.D. from Harvard University, he was awarded a Miller Research Fellowship at the University of California, Berkeley, and later a Melbourne Research Fellowship at the University of Melbourne in Australia. He enjoys backpacking, skiing and rafting with his family.


**MALE FLOWER of the surfgrass *Phyllospadix scouleri* (*left*) releases filamentous pollen that forms snowflakelike rafts. Currents carry the pollen mass to female flowers (*right*), where the pollen grains contact the stigma, resulting in fertilization.**

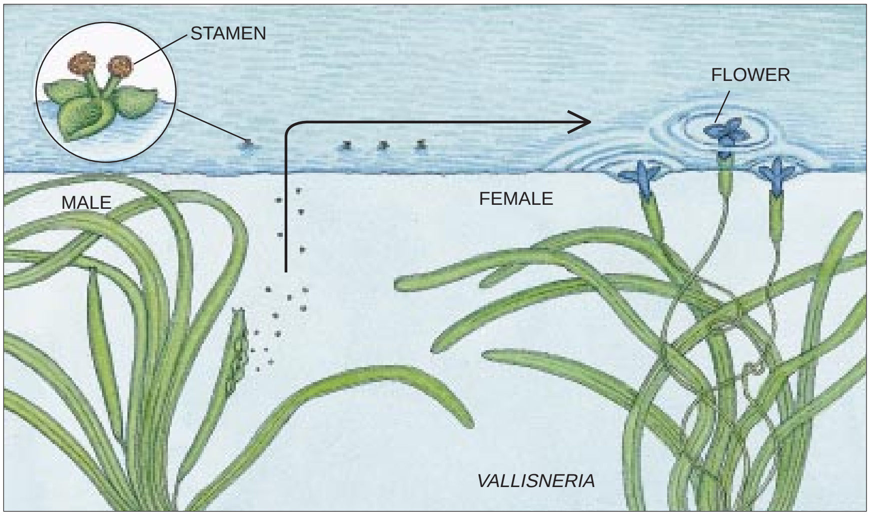


Not uncommon taxonomically or geographically, water-pollinated species occur in 31 genera in 11 different families. They range in latitude from northern Sweden to southern Argentina and in altitude from 40 meters below sea level to 4,800 meters high in the Andes, where *Elodea potamogeton* impedes navigation on Lake Titicaca on the Peru-Bolivia border. They are ecologically diverse as well, extending from tropical rain forests to seasonal desert pools. Freshwater species feed fish and wildfowl, whereas marine species support sea turtles, manatees and shellfish.

How do hydrophilous plants successfully use water as a pollination vector? Water would appear to be a poor medium in which to disperse pollen. In addition to its propensity to cause most pollen grains to burst, water can be unpredictable. Currents can be quite erratic, and tides can abruptly submerge plants or leave them far above the surface.

Historical accounts provide some information about how the plants exploited the water. Erasmus Darwin, grandfather of Charles Darwin, described the remarkable pollination system of the freshwater plant *Vallisneria* more than two centuries ago. (Darwin's description was so compelling that the Reverend William Paley cited it as evidence for the existence of God in his influential book *Natural Theology*.) Darwin found that the submerged male plants release flowers that float to the surface. Once there, the petals recurve, elevating the flower on a meniscus of water that enables it to glide with the slightest breeze. Two sterile anthers function as tiny sails, while two fertile anthers hold the spherical pollen of *Vallisneria* aloft. Each floating female flower, anchored by a long stalk to the bottom of the pond, creates a slight depression in the water's surface. As the male flower falls into the depression created by the female flower, it transfers pollen to the stigma, thus effecting pollination.

In 1792 Cavolini expanded on his earlier studies of *Zostera*, showing that another mechanism—directly releasing pollen into the water—existed for flowering aquatic plants. Cavolini also reported a second sea-grass species, *Cymodocea nodosa*, that releases its noodlelike pollen directly onto the surface of the water. He wrote that "white capillary pollen... covered for a great space the surface of the water around." Such pollen was protected from bursting because it could rapidly equilibrate with the seawater.




**ABOVE-SURFACE POLLINATION is one reproductive strategy that hydrophilous, or water-pollinated, species rely on. The pollen grains are kept above the water, where they are transported via floating male flowers that plow into female flowers to effect pollination. Examples are *Enhalus, Lagarosiphon* and *Vallisneria.***

Darwin's and Cavolini's observations received only passing interest from other investigators, who continued to concentrate on animal and wind pollination. Only in the past two decades have botanists asked broader evolutionary questions about hydrophilous pollination. Are there general principles that apply to all hydrophilous pollination systems? What features should characterize efficient systems for pollination on a two-dimensional plane such as the surface of water? What shape and size should the pollen be? Should pollen be dispersed singly or in groups?

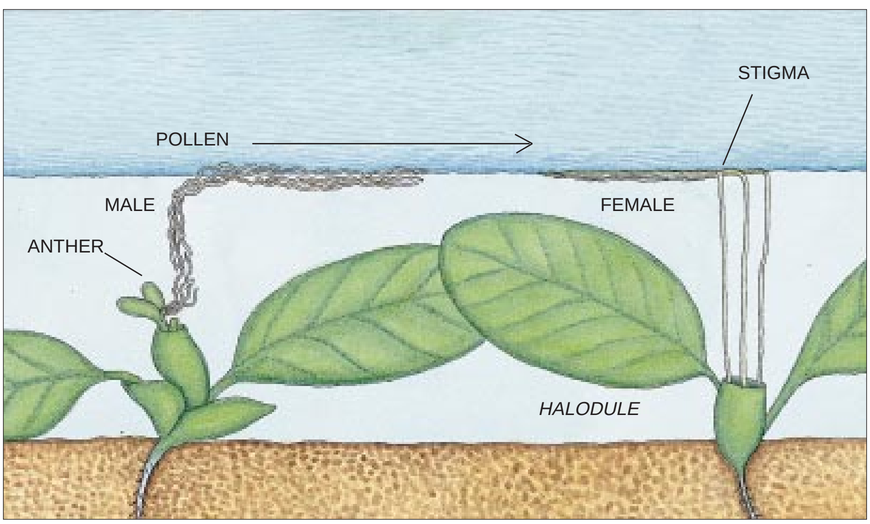
Such questions fall within the realm of search theory, first developed during World War II by the mathematician Bernard O. Koopman and his colleagues at the Antisubmarine Warfare Operations Group of the U.S. Navy. In a simple but elegant derivation, Koopman showed that if a search vehicle traces a somewhat random path in two dimensions, its probability of encountering a fixed target rapidly increases with the width of the path it traces. (Mathematically, the relation is $p = 1 - e^{-wL/A}$, where $p$ is the probability of hitting a target, $e$ is the base of natural logarithms, $w$ is the width of the path swept by the search vehicle, $L$ is the length of the path and $A$ is the search area.) Koopman's equation had immediate military significance: even a modest increase in the range of a ship-borne radar dramatically increases the probability of detecting a fixed target during a random search.

Koopman's equation applies not only to ships but also to pollen grains in the hunt for stigmas. The equation makes it possible to set a lower bound on the search efficiency of pollen grains if the grains do not trace completely random paths. If the probability of encountering a target is low (a reasonable assumption in aquatic pollination), then a small increase in the size of the search vehicle will substantially heighten the probability of hitting a stigma lying on the surface.

The evolutionary significance of Koopman's equation is clear for *Vallisneria,* the plant studied by Erasmus Darwin. Here the floating male flowers sweep a path several thousand times wider than an individual pollen grain can, thus increasing search efficiency. A similar pollination system occurs in a different species: the African pondweed *Lagarosiphon,* studied by C.D.K. Cook of the University of Zurich. As in *Vallisneria,* male flowers are released underwater and float to the surface. There three sterile stamens become erect, functioning like miniature sails. The long, fertile stamens, which unfold parallel to the water's surface, bear sticky pollen at the end of the ramrodlike filaments. The male flowers link together in large rafts, colliding with the female flowers floating on the water.

But what about such plants as the one Cavolini studied, *Z. marina,* which releases its pollen directly onto the water's surface? Are the predictions of wide search paths likewise verified? To answer these questions, R. Bruce Knox of the University of Melbourne in Australia and I examined the pollination biology of several genera of hydrophilous plants in the South Pacific. One sea grass studied was *Halodule pinifolia,* which grows along sandy coasts in the Fiji Islands. While conducting fieldwork with P. B. Tomlinson of Harvard University and John A. West of the University of California at Berkeley, I found entire *Halodule* populations to be almost completely exposed during the low spring tides. Then, the white stamens of *Halodule* slowly release the floating long, noodlelike pollen onto the surface of the sea.

The pollen noodles have coatings of proteins and carbohydrates that make them sticky. They adhere to one another, forming rafts that resemble snowflakes. Millions of these floral search ve-




**POLLEN ON WATER'S SURFACE is a second method. To form pollen rafts, some species release filamentous or oval pollen; others have spherical pollen embedded in long tubes. Stigmas tend to be threadlike in marine species such as *Amphibolis* and *Halodule* and cuplike in freshwater ones such as *Ruppia* and *Lepilaena.***

hicles are carried along as the tide returns to the shallow pools where the filamentous stigmas of the female plants float. The collision of the search vehicles with the stigmas on the water's surface results in pollination.

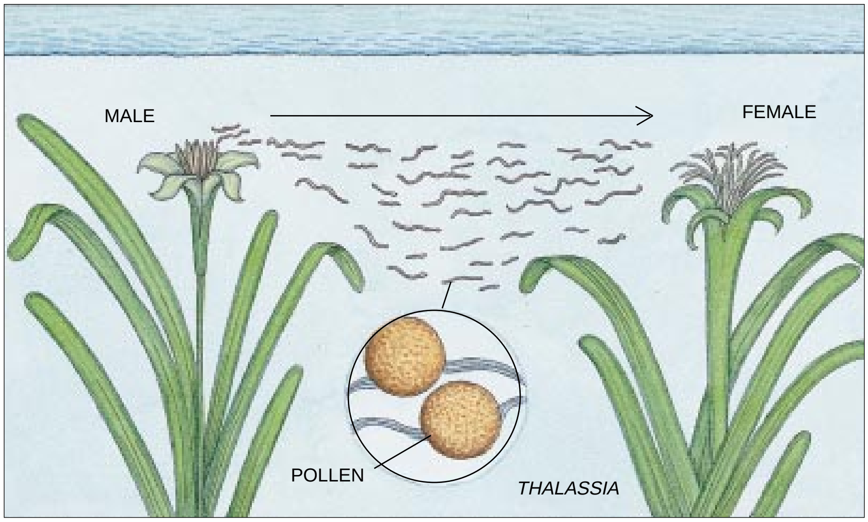
Many other sea-grass genera also assemble search vehicles that widen the search path, although the production varies slightly. In *Amphibolis* in Australia and *Thalassodendron* in Kenya, the male flowers float to the surface and there release filamentous pollen that assembles into latticelike rafts. Some unrelated sea-grass genera lack noodlelike pollen but compensate in other ways. In Fiji the small oval pollen grains of *Halophila ovalis* do not form rafts but are embedded inside sticky mucilaginous tubes that assemble into rafts, appearing as floating feathers. In Australia the freshwater species *Lepilaena cylindrocarpa* releases spherical pollen in floating mucilaginous mats that resemble tiny omelettes. These mats fall into depressions caused by the female flowers.

Thus, it is clear that many hydrophilous plants—*Vallisneria, Halodule, Halophila* and *Lepilaena,* among others—together represent an example of convergent evolution: unrelated species have evolved similar means (floating search vehicles) that serve to increase pollination efficiency. Of course, some hydrophilous species have unique adaptations, mechanisms that perhaps are indicative of evolutionary whimsy. What military strategist could fail to be intrigued by the horned pondweed *Zannichellia palustris?* Cook and I and our respective students You Hao Guo and Rebecca Sperry found *Zannichellia* to form ominous floating clouds that slowly drop round pollen grains onto the stigmas below. Fortunately, in this case, being hit by such a botanical depth charge results not in flower destruction but plant reproduction. And what golfer could fail to be delighted by the pollination system of another freshwater plant, *Hydrilla verticillata,* also studied by Cook? Here, male flowers float to the water surface and explosively catapult their pollen through the air to the floating female flowers. Because *Hydrilla* pollen is destroyed by contact with water, only the holes-in-one count. Given the vast amount of pollen released, the probability that a stigma is struck is actually fairly high.

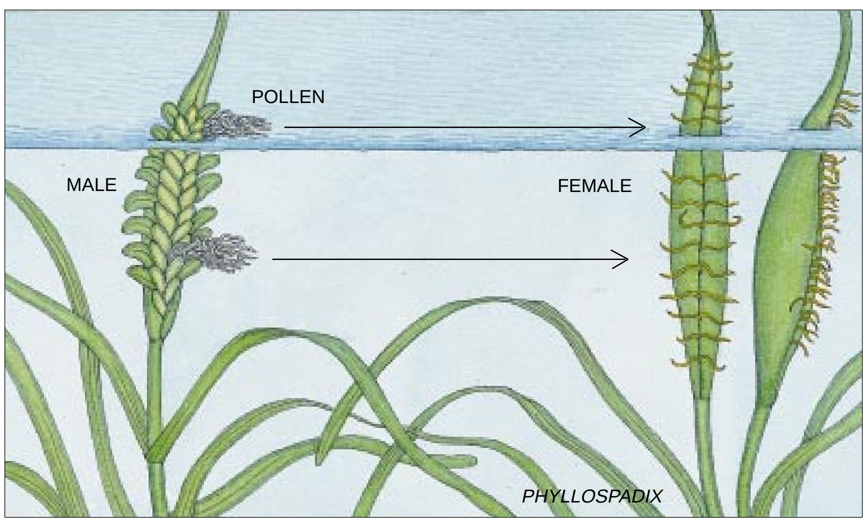
So far I have discussed the pollination systems of plants whose pollen is transported above or on the surface of the water. In such cases, the pollen moves in two dimensions. Some species have pollination systems that operate in three dimensions—that is, below the surface [see "Submarine Pollination," by John Pettitt, Sophie Ducker and Bruce Knox; SCIENTIFIC AMERICAN, March 1981].

Pollination strategies in three dimensions seem to be less common than the two-dimensional varieties, probably because they are intrinsically less efficient. In three dimensions, pollen and stigmas are dispersed throughout a volume rather than concentrated on a single plane (such as the water's surface), and so more pollen is needed. The consequences of random motion also differ greatly in three dimensions. A random hunt is far more efficient in two dimensions because the search path is recurrent: any stigma on the plane will eventually be hit given enough time. The same cannot be said for a three-dimensional search, even given an infinite amount of time [see "Brownian Motion and Potential Theory," by R. Hersch and R. J. Griego; SCIENTIFIC AMERICAN, March 1969].

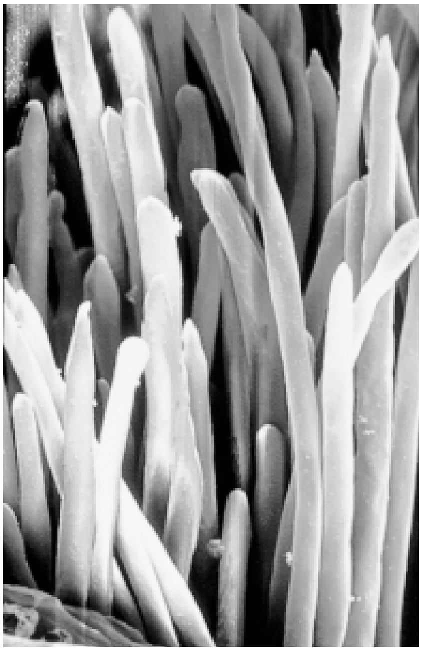
Three-dimensional pollination has tended to evolve in hydrophilous plants that ordinarily would have little opportunity to engage in two-dimensional pollination—that is, in plants that are always submerged. An example is the tur-

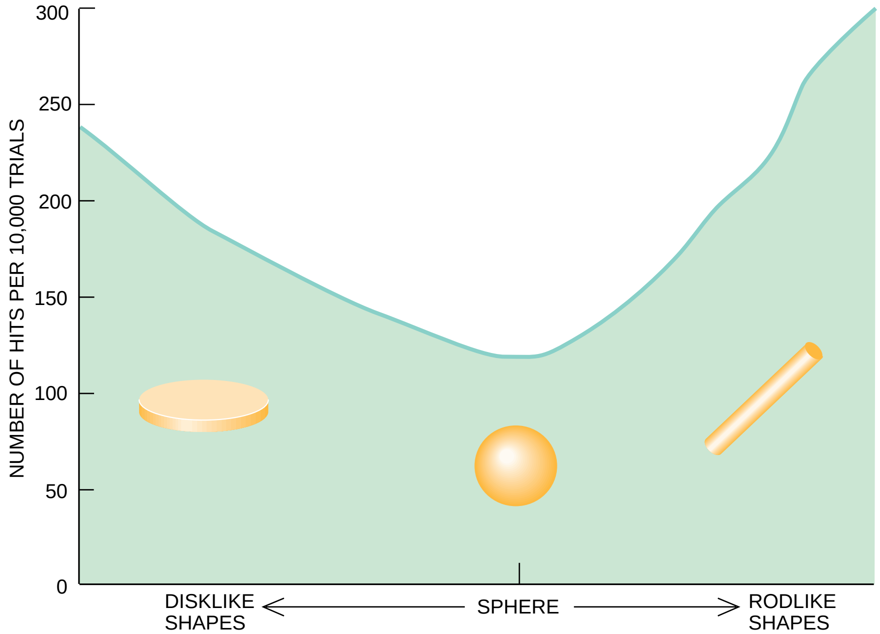



**UNDERWATER POLLINATION is a third method of reproduction. In some species, filamentous pollen or pollen grains in mucilaginous strands are carried underwater by currents. In others, pollen denser than water is released. Examples include *Thalassia, Syringodium* and a species of the freshwater genus *Lepilaena.***




**MIXED MODE of pollen transport, used by *Phyllospadix* and *Zostera,* combines surface with submarine pollination. Surface pollen tends to be slightly more filamentous than submarine pollen. Underwater delivery occurs in subtidal zones—that is, in populations that never reach the surface, even during low tide.**






**OPTIMAL POLLEN SHAPE for underwater pollination was determined by analysis on a supercomputer. Spheres were deformed into ellipses of equal volume, either as long rods or as disks. The rods proved best at striking a target and compare well with actual submarine pollen grains, such as those of *Phyllospadix* (*photograph*).**

tle grass *Thalassia testudinum,* a Caribbean plant that Tomlinson, Thomas Elmqvist of Umeå University in Sweden and I studied on the island of St. Croix. *Thalassia* releases its round pollen underwater, embedded in elongated strands of mucilage. Having been driven underwater by waves, the strands collide there with the stiff, bristly stigmas of the female plants. A similar underwater pollination mechanism of embedding round pollen on long, mucilaginous strands is a feature of an unrelated freshwater species: *Lepilaena bilocularis,* found at Lake Ellesmere on South Island of New Zealand.

I wanted to see if this pearls-on-a-string approach by these geographically remote and taxonomically distant species represented an efficient three-dimensional strategy. Together with my mathematics students Tyler Jarvis, now at Princeton University, and Scott Cromar, now at Rutgers University, I ran a series of numerical experiments on a supercomputer at the IBM Scientific Computing Center in Palo Alto, Calif. We deformed spheres elliptically to varying degrees while keeping their volume constant, thus producing a variety of long strings or flattened Frisbee-like disks. We then determined how frequently each shape encountered randomly positioned targets. In our simulations we found that all the elliptical deformations were more efficient in three-dimensional searches than were the spheres [*see illustration above*].

Recently my student David Smith, now at Princeton, David A. Olson, then at New York University's Courant Institute of Mathematical Sciences, and I confirmed these numerical experiments analytically. The equations also revealed something the supercomputer experiments had missed: shapes that are only slightly disklike are the least efficient at three-dimensional search. Still, the general conclusion holds. The search vehicles of *T. testudinum* in the Caribbean Ocean and *L. bilocularis* in New Zealand—each consisting of pollen attached to a mucilaginous strand—represent highly efficient solutions to the problems of locating objects scattered in three dimensions.

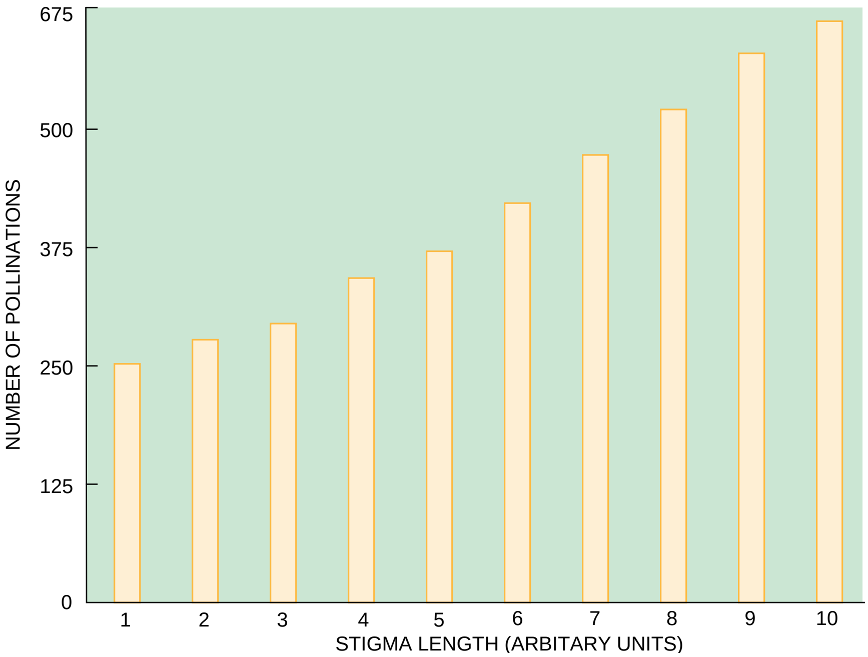
Some hydrophilous species mix both two- and three-dimensional strategies. For example, *Enhalus acoroides,* a sea grass I studied in the Indonesian island of Banda, flowers only during low spring tides. Like *Vallisneria,* the male flower buds float to the water's surface, open and are captured within the rumpled petals of the female flowers. As the tide rises, the petals of the female flowers close, trapping the male flowers within. But *Enhalus* disperses some pollen directly on the water's surface.

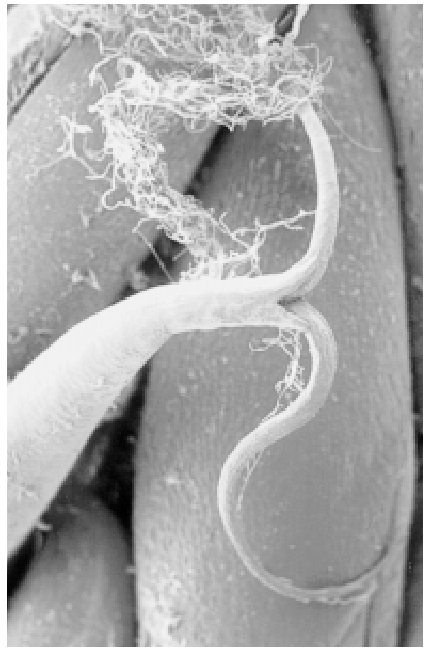
The surfgrass *Phyllospadix scouleri* also disperses noodlelike pollen both on and below the surface. Studying a population along the northern California coast, Tomlinson, my student Kevin Nieznanski and I observed floating pollen search vehicles of *Phyllospadix* collide with female stigmas. But electron micrographs of stigmas collected far below the water's surface also revealed pollen, indicating that submarine pollination can take place.

Yet another example of mixed surface and submarine pollination occurs in the plant originally studied by Cavolini: *Z. marina.* A.W.A.M. de Cock of Catholic University in the Netherlands found that although *Zostera* pollen released underwater in aquariums slowly sinks, pollen dispersed at low tide in intertidal populations floats on the surface. Roger H. Laushman of Oberlin College, Mary H. Ruckelshaus of the University of Washington and I, working at Friday Harbor Marine Laboratories in the San Juan Islands, off Washington State, confirmed de Cock's report. We found that pollen rafts resembling snowflakes float on the surface; however, we also observed linear bundles of noodle-shaped pollen dispersed beneath the surface of the sea.

Efficient search methods are not the only way to increase the chances of pollination. Evolution can also select for stigmas that can readily be found. Unfortunately, the design of targets that could easily be hit was not considered a high-priority task by the military strategists who developed search theory. So, much of the analysis of target design has fallen to botanists.

From the stigma's point of view, there are at least two possible ways to heighten the chances of being struck by a pollen grain or pollen raft. First,



**OPTIMAL STIGMA SHAPE in surface pollination was determined by deforming spheres into increasingly longer ellipses. Pollen grains collided with the long, elliptical stigmas more frequently than they did with the round stigmas. The results compare favorably with stigmas of real plants, such as that of *Zostera marina* (*photograph, with pollen strands clinging*).**

the effective area of the target can be broadened. Second, the stigma's shape can be altered to increase the probability of encounter.

The floral structures of many water-pollinated plants embody the first solution. Some blossoms, such as those of *Vallisneria* and *Enhalus,* create depressions in the water's surface, forming a target area much larger than a stigma or even an entire female flower. Another example is provided by *Lepilaena cylindrocarpa* in Australia. Usher Posluszny of the University of Guelph pointed out to me a small bract that holds three separate stigmas together before fertilization. Thus united, the three stigmas of the female flower, each resembling a baseball mitt, cause a dip in the surface of the water.

The oscillatory motion of some plant parts can also enlarge the effective target area. In a slow-moving stream, the elongated stalk of *Ruppia marina* functions much like a windshield wiper: it causes the flowers to sweep gently back and forth, collecting the snowflake-shaped pollen rafts. Floral structures may also have localized effects. Joseph D. Ackerman, then at Cornell University, observed how submerged flower clusters of *Zostera* affect underwater flow patterns. He found that under laboratory conditions the altered flow can concentrate pollen near the stigmas.

Investigation into the second possible means of increasing target efficiency—changing target shapes—has begun only in the past few years. In a series of supercomputer simulations, James A. Sethian of Berkeley and I found that in two dimensions, filamentous stigmas are far more likely to be hit than are circular stigmas. The extremely long, thread-like stigmas of the sea-grass genera *Halodule, Halophila* and *Thalassodendron* seem to confirm the prediction [*see illustration above*].

Field observations also contradict the assumptions made by some previous workers, who thought that hydrophilous plants rarely flower or pollinate. I have found that flowering and cross-pollination (the transfer of pollen from one individual plant to another) to be common, though sometimes ephemeral, events. The frequency of flowering—and by implication cross-fertilization—can directly affect genetic diversity of the populations. By using a technique called starch-gel electrophoresis, Laushman has found cross-pollination in *Zostera* populations to be relatively frequent, particularly in the intertidal area where surface pollination happens.

Yet other investigators find little diversity. Based on his studies of the freshwater species *Ceratophyllum demersum,* Donald H. Les of the University of Wisconsin at Milwaukee believes hydrophilous plants may in large part be genetically uniform. Such uniformity could arise from self-pollination. For instance, C. Thomas Philbrick of Rancho Santa Ana Botanic Garden in Claremont, Calif., has found that in a submerged species of *Potamogeton,* minute air bubbles can carry pollen from the anther to the stigma of the same flower. Genetic homogeneity could also result from infrequent flowering and large clone sizes in a population.

Although the mechanics of water pollination has not been able to detail the genetics of hydrophilous populations, it does offer clues to explain why such plants followed an evolutionary path different from their terrestrial counterparts. The noodlelike shape of the pollen is one distinctive feature of hydrophilous plants. To study the emergence of such pollen, C. J. Humphries of the London Natural History Museum and I have attempted to reconstruct the family tree of the sea-grass family Cymodoceaceae. We found that this family, which includes *Thalassodendron, Amphibolis, Syringodium, Halodule* and *Cymodocea,* shares a common ancestor with *Posidonia* (a sea grass found in the Mediterranean and Western Australia) and Zosteraceae, a family that encompasses *Zostera, Phyllospadix* and the Australian genus *Heterozostera.* Our analysis suggests that this common ancestor had noodle-shaped pollen. Thus, noodlelike pollen probably evolved only once, explaining why other unrelated sea-grass genera such as *Halophila, Thalassia* and *Enhalus* still retain round pollen similar to that of terrestrial plants.

Yet another striking feature of hydrophilous plants has to do with dioecism—the separation of male and female flowers into distinct plants. More

**BREACHING THE WATER'S SURFACE during the spring low tides is the only time many marine hydrophilous plants, such as a population of *Phyllospadix* along the coast of Monterey, Calif. (*left*), have to release pollen or flowers on the surface. The pollen of many such hydrophilous species assemble into rafts, such as those of *Amphibolis* (*right*).**

than half of all hydrophilous species are dioecious. In contrast, only about 3 or 4 percent of terrestrial species are dioecious. Dioecism in hydrophilous plants might have evolved because of the physical incompatibility between the mechanisms of waterborne pollen dispersal and pollen capture within a single flower. A combined floral morphology would likely result in a flower capturing its own pollen before release, defeating the ability of the plants to cross-fertilize.

Another possible reason for the observed high rate of dioecism is that it helps to prevent inbreeding. Because many hydrophilous populations consist of large clones, inbreeding can result if male and female parts occur on the same plant. Alternative mechanisms of outbreeding typical of terrestrial plants probably would not function as well in hydrophilous plants. For example, the maturation of male and female flowers at different times, a phenomenon typical in land plants, would probably be difficult to synchronize in a large sea-grass clone.

The unusual features of hydrophilous plants render them not only fascinating but also extremely useful as study organisms for many problems in the evolutionary ecology and demography of plants. The life span of hydrophilous pollen is easy to determine because the pollen grains of many hydrophilous plants display cytoplasmic streaming. This feature, readily observed under a microscope, is simply the movement of cytoplasm within the pollen grain. I found that *Zostera* pollen, for example, has a strikingly short lifetime: almost all the pollen dies within eight hours after release. In contrast, determining the life span of terrestrial pollen is much more difficult: terrestrial pollen generally must be stained to determine viability.

Furthermore, the fate of individual pollen grains of surface-pollinated species is easily determined. One can simply examine, frame by frame, videotapes of the water surface and trace the paths of the grains. This approach is clearly impractical with plants whose pollen is carried by wind or animals.

Hydrophilous species may yield insights into patterns of geographic colonization and the genetics of plant populations because the populations of freshwater hydrophilous plants are in essence reproductively isolated. Waterborne pollen cannot move between ponds. Gene flow between populations must therefore transpire primarily by transfer of fruits or small vegetative pieces that break away. Vigorous clonal growth in most water-pollinated plants also makes them ideal for ecological experiments: one can expose the same genotype to many different environments.

Compared with the widespread interest in wind and animal pollination, little work has been done on hydrophilous pollination since Cavolini first discovered *Zostera* flowers in the Bay of Naples. Although water-pollinated plants do not produce large, showy flowers, as do many terrestrial plants, I find them to be exquisitely beautiful. The plants provide opportunities for both amateur and professional botanists to make original observations in natural history and to study more closely convergent evolution in aquatic environments.

FURTHER READING

SEARCH AND SCREENING. B. O. Koopman. Pergamon Press, 1980.

HYDROPHILOUS POLLINATION. P. A. Cox in *Annual Review of Ecology and Systematics,* Vol. 19, pages 261–280; 1988.

TWO-DIMENSIONAL POLLINATION IN HYDROPHILOUS PLANTS: CONVERGENT EVOLUTION IN THE GENERA HALODULE (CYMODOCEACEAE), HALOPHILA (HYDROCHARITACEAE), RUPPIA (RUPPIACEAE), AND LEPILAENA (ZANNICHELLIACEAE). P. A. Cox and R. B. Knox in *American Journal of Botany,* Vol. 76, No. 2, pages 164–175; February 1989.

SURFACE AND SUBMARINE POLLINATION IN THE SEAGRASS *ZOSTERA MARINA L.* Paul A. Cox, Roger H. Laushman and Mary H. Ruckelshaus in *Botanical Journal of the Linnaean Society,* Vol. 109, No. 2, pages 281–291; June 1991.

UNDERWATER POLLINATION, THREE-DIMENSIONAL SEARCH, AND POLLEN MORPHOLOGY: PREDICTIONS FROM A SUPERCOMPUTER ANALYSIS. P. A. Cox, S. Cromar and T. Jarvis in *Pollen and Spores: Patterns of Diversification.* Edited by S. Blackmore and S. H. Barnes. Oxford University Press, 1992.

# Simulating Brain Damage

*Adults with brain damage make some bizarre errors when reading words. If a network of simulated neurons is trained to read and then is damaged, it produces strikingly similar behavior*

by Geoffrey E. Hinton, David C. Plaut and Tim Shallice

In 1944 a young soldier suffered a bullet wound to the head. He survived the war with a strange disability: although he could read and comprehend some words with ease, many others gave him trouble. He read the word *antique* as "vase" and *uncle* as "nephew."

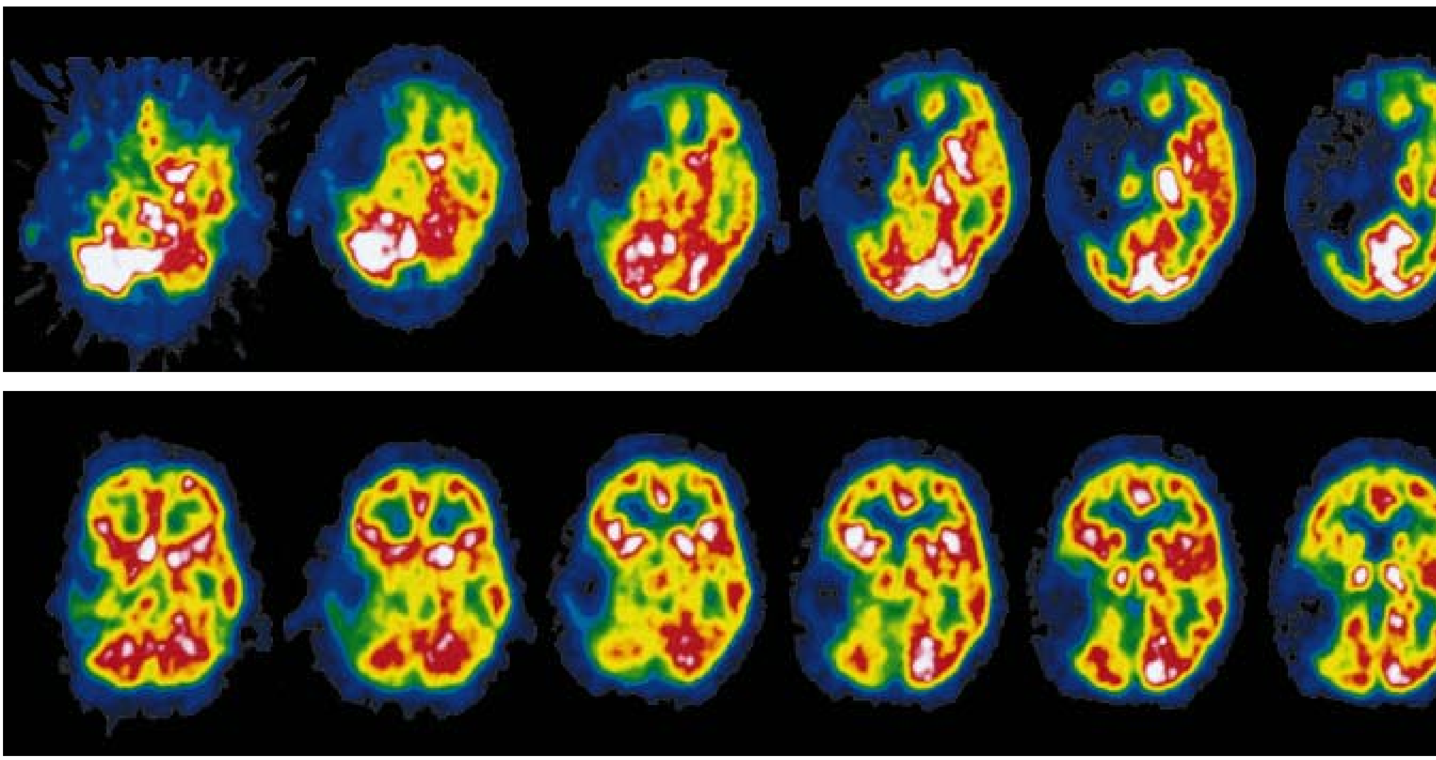
The injury was devastating to the patient, G.R., but it provided invaluable information to researchers investigating the mechanisms by which the brain comprehends written language. A properly functioning system for converting letters on a page to spoken sounds reveals little of its inner structure, but when that system is disrupted, the peculiar pattern of the resulting dysfunction may offer essential clues to the original, undamaged architecture.

During the past few years, computer simulations of brain function have advanced to the point where they can be used to model information-processing pathways. We have found that deliberate damage to artificial systems can mimic the symptoms displayed by people who have sustained brain injury. Indeed, building a model that makes the same errors as brain-injured people do gives us confidence that we are on the right track in trying to understand how the brain works.

We have yet to make computer models that exhibit even a tiny fraction of the capabilities of the human brain. Nevertheless, our results so far have produced unexpected insights into the way the brain transforms a string of letter shapes into the meaning of a word.

When John C. Marshall and Freda Newcombe of the University of Oxford analyzed G.R.'s residual problems in 1966, they found a highly idiosyncratic pattern of reading deficits. In addition to his many semantic errors, G.R. made visual ones, reading *stock* as "shock" and *crowd* as "crown." Many of his misreadings resembled the correct word in both form and meaning; for example, he saw *wise* and said "wisdom."

Detailed testing showed that G.R. could read concrete words, such as *table,* much more easily than abstract words, such as *truth.* He was fair at reading nouns (46 percent correct), worse at adjectives (16 percent), still worse at verbs (6 percent) and worst of all at



**BRAIN IMAGES show damage to the language-processing areas of patients with acquired dyslexia, which can now be modeled by artificial neural networks. (These positron-emission tomography scans, made by Cathy J. Price and her colleagues at the MRC Cyclotron Unit in London, measure activity of the brain in successive horizontal slices, starting at the top. Low**

function words, such as *of* (2 percent). Finally, he found it impossible to read wordlike nonsense letter strings, such as *mave* or *nust.*

Since then, clinicians have studied more than 50 other patients who make semantic errors in reading aloud, and virtually all of them show the same strange combination of symptoms. In 1973 Marshall and Newcombe described two contrasting types of acquired dyslexia. So-called surface dyslexics misread words that are pronounced in an unusual way, often giving the more regular pronunciation; a surface dyslexic might read *yacht* as "yatched." In contrast, a "deep" dyslexic patient like G.R. might read *yacht* as "boat."

To explain the existence of these two types of dyslexia, Marshall and Newcombe proposed that the information processed in normal reading travels along two distinct, complementary routes. Surface dyslexics retain the phonological route, which relies on common spelling-to-sound correspondences. Deep dyslexics, meanwhile, retain the semantic route, which allows the meaning of a word to be derived directly from its visual form (when it can be derived at all). A person reading words aloud via the semantic route derives pronunciation entirely from meaning.

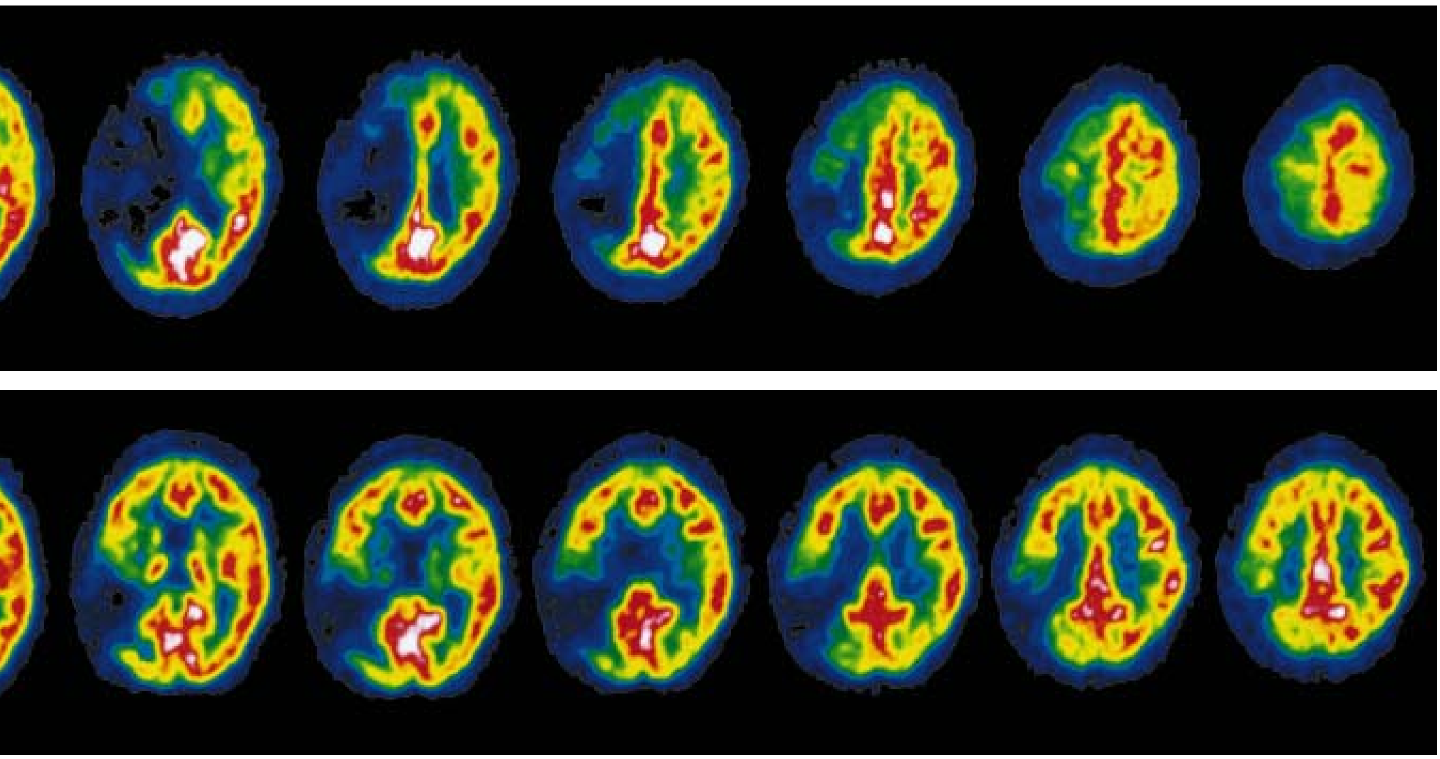
According to Marshall and Newcombe, the errors produced by deep dyslexics reflect how the semantic route operates in isolation. Later empirical findings suggest that this account is oversimplified, but the notion of a semantic route is still generally accepted. It now seems likely that deep dyslexics not only lose their phonological route but have damage somewhere along the semantic one as well.

The hypothesis that reading depends on multiple routes that can be separately damaged has proved fruitful in classifying patients but less useful in understanding the precise nature of their injuries. Max Coltheart of Macquarie University in Australia and Eleanor M. Saffran of Temple University have both proposed, for example, that the reading of deep dyslexics may bear a strong resemblance to that of patients who have only the right hemisphere of their brain functioning.

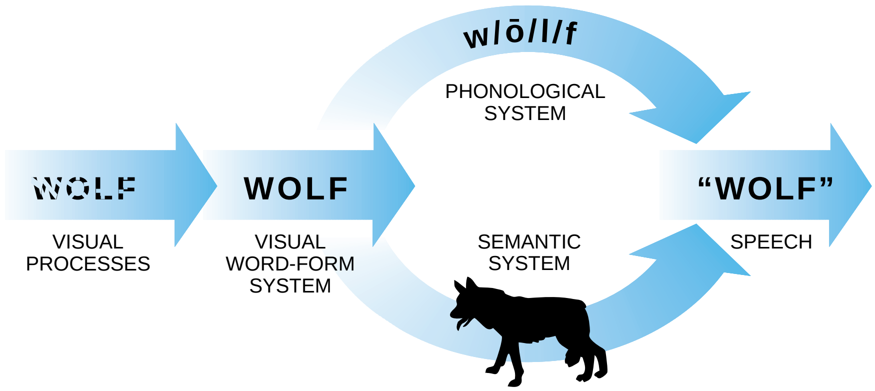
This explanation, however, provides little insight into the highly characteristic pattern of errors that typically occurs in acquired dyslexia. Any detailed explanation of how errors arise and why they form consistent patterns requires a model of how that information is processed in each route—and of how this processing goes wrong when the neural circuitry is damaged. Psychologists often use abstract, algorithmic descriptions of the way that the brain handles information. These descriptions obviously cannot be subjected to the kinds of injuries that brain cells may incur.

As a result, we have turned to neural networks—idealized computer simulations of ensembles of neurons. We have developed networks that perform the role of the semantic route, and then we have selectively removed connections between neurons to see how their be-


GEOFFREY HINTON, DAVID PLAUT and TIM SHALLICE use artificial neural networks to investigate the behavior of the brain. Hinton is the Noranda Fellow of the Canadian Institute for Advanced Research and professor of computer science and psychology at the University of Toronto. He has been studying representation and learning in neural networks for more than 20 years. Plaut is a postdoctoral research associate in the psychology department at Carnegie Mellon University, where he earned his doctorate in computer science in 1991. Shallice is professor of psychology at University College, London, where he received his doctorate in 1965. His research has mainly focused on what can be understood about the normal cognitive system by studying impairments resulting from neurological disease, with occasional forays into the organization of short-term memory and the nature of will and consciousness.

**levels of activity appear in blue and high levels in white.) One patient (*top row*) has lost almost all function in the left hemisphere of the cerebral cortex, except for the most posterior regions. The other has sustained damage to the parietal and temporal lobes of the left hemisphere, regions generally believed to be crucial for processing language.**




**TWO PATHWAYS in the brain are responsible for the mental processing and pronunciation of written words. One (the phonological route) derives pronunciation from spelling, the other (the semantic route) from meaning. Deep dyslexics have lost the phonological route completely and have suffered damage to the semantic route as well.**

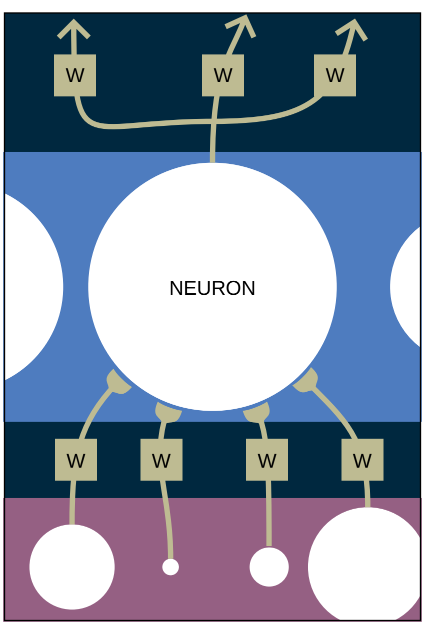
havior changes. A few years ago we designed a simple network to mimic the semantic route and found that damaging any part of it could reproduce several of the symptoms of deep dyslexia. We have since made more detailed models to learn which aspects of neural-network architectures were responsible for this behavior. We have also extended the approach to account for additional symptoms of deep dyslexia.

Our models of the semantic route consist of interconnected units representing neurons. Each neuron unit has an activity level (between 0 and 1) that depends on the inputs it receives from other neurons. Connections between units have an adjustable weight that specifies the extent to which the output of one unit will be reflected in the activity of the unit it is feeding. These weights, along with the pattern of connections among neurons, determine the computation that the network performs.

The first version of our network consisted of three sets of units: “grapheme” units, each of which represented a particular letter in a specific position within the word; “sememe” units that represented the meanings of words; and a layer of intermediate units that make it possible to learn complex associations. A completely general network would require 26 grapheme units for each position within a word, but we used a simplified vocabulary that permitted a smaller number. The grapheme units in the first position were all consonants, for instance, and those in the second, all vowels.

The sememe units do not correspond directly to individual word meanings but rather to semantic features that describe the thing in question. The word *cat* activates such units as “mammal,” “has legs,” “soft” and “fierce.” Units representing such semantic features as “transparent,” “tastes strong,” “part of limb” or “made of wood” remain quiescent. Our network has 68 sememe units representing both physical and functional attributes of a word’s definition. Each word that we chose was represented by a different combination of active and inactive sememe units.

To make our neural network produce the correct pattern of semantic features for each word, we had to set the weight on each connection to the appropriate value. These weights are set not by hand but rather through a learning procedure—an algorithm for programming neural networks. To teach a network a task, one starts with random weights and then presents the network repeatedly with a “training set” of input patterns (in this case, letters in specified positions). The algorithm adjusts the weights after each training run to reduce the difference between the network’s output and the “correct” response.




**IDEALIZED NEURON is the basis for artificial neural networks. It sums the weighted inputs that it receives from other neurons (*bottom*) and generates an activation level between 0 and 1. It then passes this activation (through weighted connections) to other neurons. The set of weights and connections in a neural network determines its behavior.**

Neural-net workers have known since the 1950s how to adjust weights in simple, two-layer networks, but training networks with a greater number of layers is more difficult. In particular, it is not immediately obvious how to set the weights on the connections from the input units to the intermediate units because there is no way to determine, a priori, which intermediate units should be active for any given input and output.

During the 1980s, however, neural-net researchers developed a number of different methods for training multilayer networks. These methods apportion changes to the connection weights of each layer according to their contribution to the error. Over the course of many training cycles, the resulting weights converge to yield a network that produces the correct results. Depending on the initial random weights, learning may result in any of a number of sets of weights, each of which leads the network to produce correct answers for its training inputs. (For further details of the learning procedure, see “How Neural Networks Learn from Experience,” by Geoffrey E. Hinton; SCIENTIFIC AMERICAN, September 1992.)

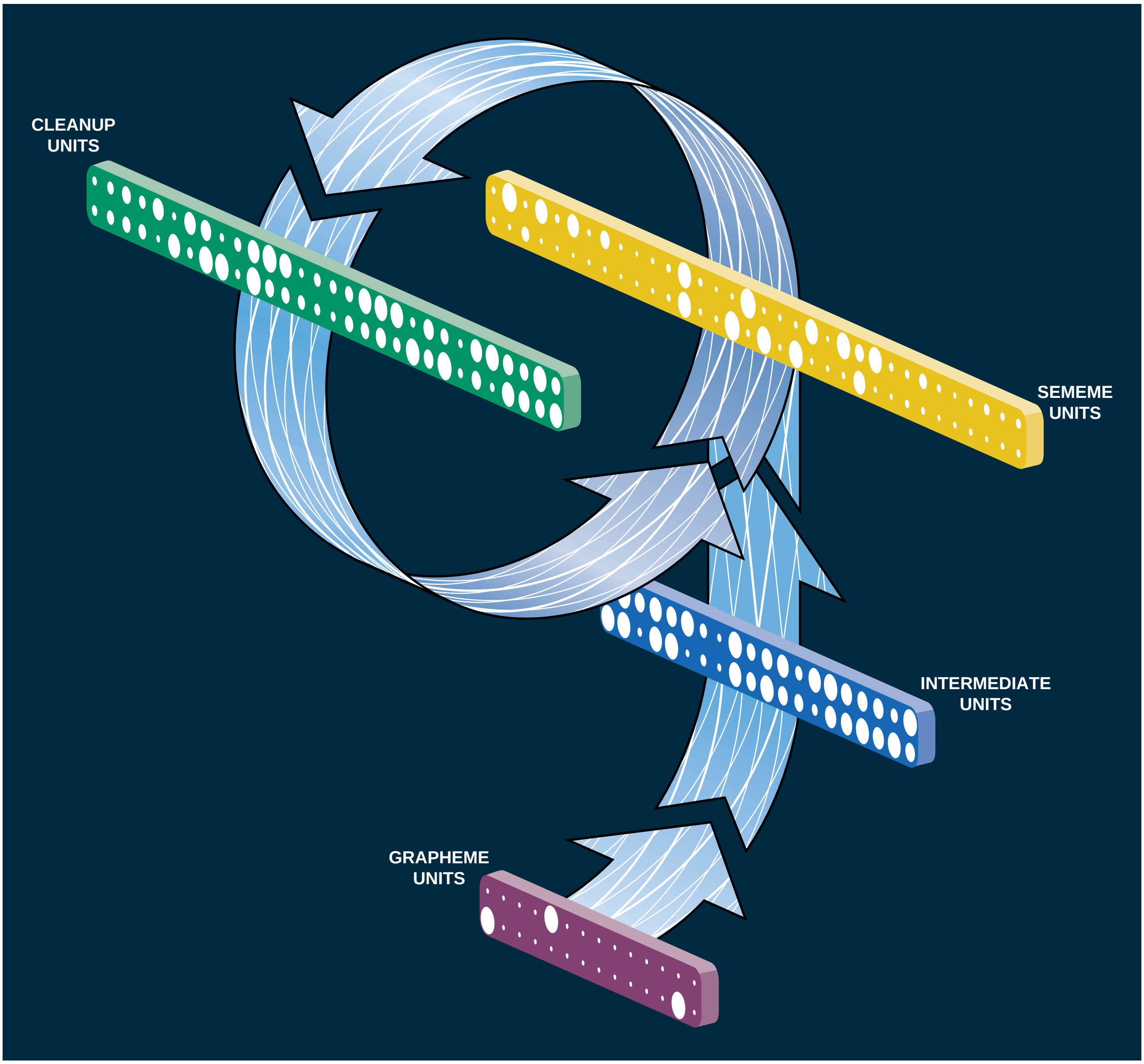
In theory, these learning procedures can get stuck in so-called local minima—configurations of weights that are incorrect but for which any small change would only make the network’s errors worse. In practice, however, a network almost always learns nearly optimal solutions. In addition, some of the learning procedures are more biologically plausible than others, but our results do not seem to depend on which method we use. We suspect that even if the brain uses a quite different learning procedure, the resulting neural circuitry will still resemble the structure that our network develops. Thus, our explanation of what happens when the network is damaged may be correct even if its learning procedures are not.

Although our initial network, with one intermediate layer, could learn to map word-forms to their semantic features, it was not really satisfactory. It had a strong tendency to map very similar inputs (such as *cat* and *cot*) to similar outputs unless subjected to excessively long training. We addressed this problem by adding another layer of "cleanup" neurons. If the original set of connections produces a sloppy answer, the new units will change it to produce exactly the correct semantics. The number of word meanings is limited, so the pathway from the input need only get the activities of the sememe units closer to the correct meaning than to any other. The same learning techniques that succeed on networks with a single intermediate layer can direct the learning of nets containing multiple intermediate layers or even networks whose units are connected in cyclical fashion.

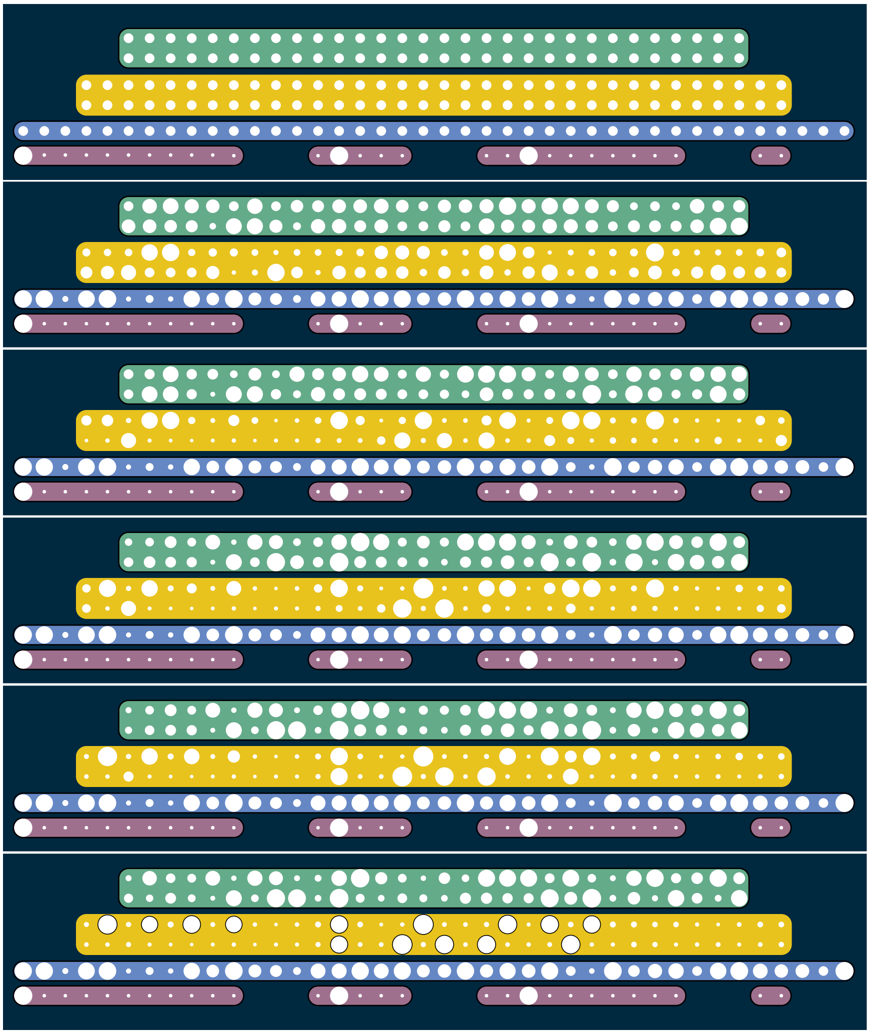
The most natural way to implement this cleanup mechanism is with a feedback loop. The output of the sememe units goes to the cleanup units, and their output goes to the inputs of the sememe units. Each time activity flows around the loop, the influence of the cleanup units on the sememe units (and vice versa) will yield a pattern of semantic features that is closer to the correct one.

The feedback loop introduces a new characteristic into the behavior of our neural network. The original network was static—any given input would cause the network to produce a corresponding output pattern, and that pattern did not change as long as the input stayed constant. The output of the new network, however, is dynamic; it settles gradually into a stable pattern.

Consequently, we have found it useful to think of the network's output not just as a list of active semantic features but rather as motion through a multidimensional "semantic space," whose coordinates are defined by all the semantic features that the network can represent. Every point in the space corresponds to a specific pattern of activity among the




**NEURAL NETWORK FOR READING contains four layers. The first responds to the letters in each word. Connections between input and intermediate units and between intermediate and "sememe" units convert the word-form to a representation in terms of semantic features, such as size, edibility or aliveness. "Cleanup" units are connected to sememe units in a feedback loop that adjusts the sememe output to match the meanings of words precisely.**




**ACTIVATION LEVELS of neurons in the network change with time as the net processes the word *bed*. At first, many of the sememe units are activated to varying degrees, but interaction with the cleanup units strengthens the activation of some and weakens that of others until the output converges.**

sememe units, but only a few of those patterns correspond to valid meanings. The correct meanings of words are points in semantic space.

The first three layers of the network, seen according to this perspective, take a word-form and convert it to a position somewhere in semantic space. Activity in the cleanup layer then draws the output of the network to the point corresponding to the closest meaning. The region around each word is what physicists and mathematicians know as a point attractor—whenever the network's initial output appears within a certain region, the network's state will inexorably be drawn to one position within the region.

This notion of a semantic space dotted with attractors representing the meanings of words has proved valuable for understanding how our network operates and how it can make the same semantic errors that dyslexics do. If we damage the network by randomly changing the weights in the cleanup mechanism, for example, the boundaries of the attractor for each word will change. As a result, if the network is in a region in semantic space where it was previously drawn to one word, it may now be drawn to a semantically related one instead. Alternatively, if we disrupt the pathway coming from the input, the network's initial output may be closer to the meaning of a semantically related word than to the meaning of the word originally presented.

This result clears up one of the first puzzles presented by deep dyslexia: why damage to any part of the brain's semantic route produces an essentially similar pattern of misreadings. Neurologists and others had wondered how damage near the input—the visual part of the reading system—could cause semantic errors. According to our models, these errors arise naturally as the cleanup neurons use semantic information to try to make sense of the output of the damaged earlier stages.

The notion of attractors helps to explain another anomaly in the data as well. Almost all patients who make semantic errors also make some visual errors—they confuse a word like *cat* with a visually similar word like *cot*. They do not, however, make the sounding-out errors of surface dyslexics ("loave" for *love* or "deef" for *deaf*). This invariable connection between semantic errors and visual errors is odd. Some patients must have damage solely to the later stages of their processing systems, and one would intuitively expect them to make only semantic errors.

After implementing our neural-network model, we discovered to our great surprise that damage to the semantic cleanup circuit sometimes caused visual errors. Retrospectively, we can understand why: the earlier layers of an undamaged network can afford to produce somewhat similar semantic outputs for the words *cat* and *cot* because the cleanup circuit will steer each to its proper meaning. But when the cleanup circuit is damaged and the shapes of each attractor change, the output of the sememe units may fall into the attractor for a visually similar but semantically unrelated word.

This explanation did not initially occur to us because it relies on the idea that the boundary of the attractor for *cat* can come very close to the one for *cot* even though the two words are semantically dissimilar. One would expect the attractors for many other meanings to come between those for *cat* and *cot*. In a two-dimensional space this intuition is correct: if we choose 40 points at random to represent word meanings and construct fairly compact attractors around each point, the attractors for dissimilar meanings will not come anywhere near one another.

It is very dangerous, however, to assume that the same is true in spaces that have many dimensions. Our network represents 68 semantic features in its sememe units, and so the attractors for each of its 40 words reside in a 68-dimensional space. It turns out that in 68 dimensions, the midpoint between any two randomly chosen points is almost certainly closer to each of those points than it is to any of 38 other random points. Consequently, the attractors for *cat* and *cot* can have a common border without any other attractors getting in the way. Avoiding obstacles is easy in 68-dimensional space.

Although our network was able to mimic both the correct and dysfunctional mapping of word-forms to meanings, that does not mean its architecture is the only possible one for the brain's semantic processing route. To determine the range of possible alternatives, we investigated the effects of damage on several different architectures, each designed to evaluate one aspect of the original network design.

We programmed versions of the neural network that contained connections among the sememe units and ones that lacked such connections; we also programmed some networks so that each neuron in one layer was connected to every neuron in the succeeding layer and others whose connections were sparse. In addition, we moved the cleanup units so that they performed their work ahead of the sememe units, and we combined the cleanup units with the intermediate layer. We even changed the arrangement of neurons in the input layer to alter the way that words were represented and added an output network that converted meanings to strings of phonemes, so that the system actually spoke.

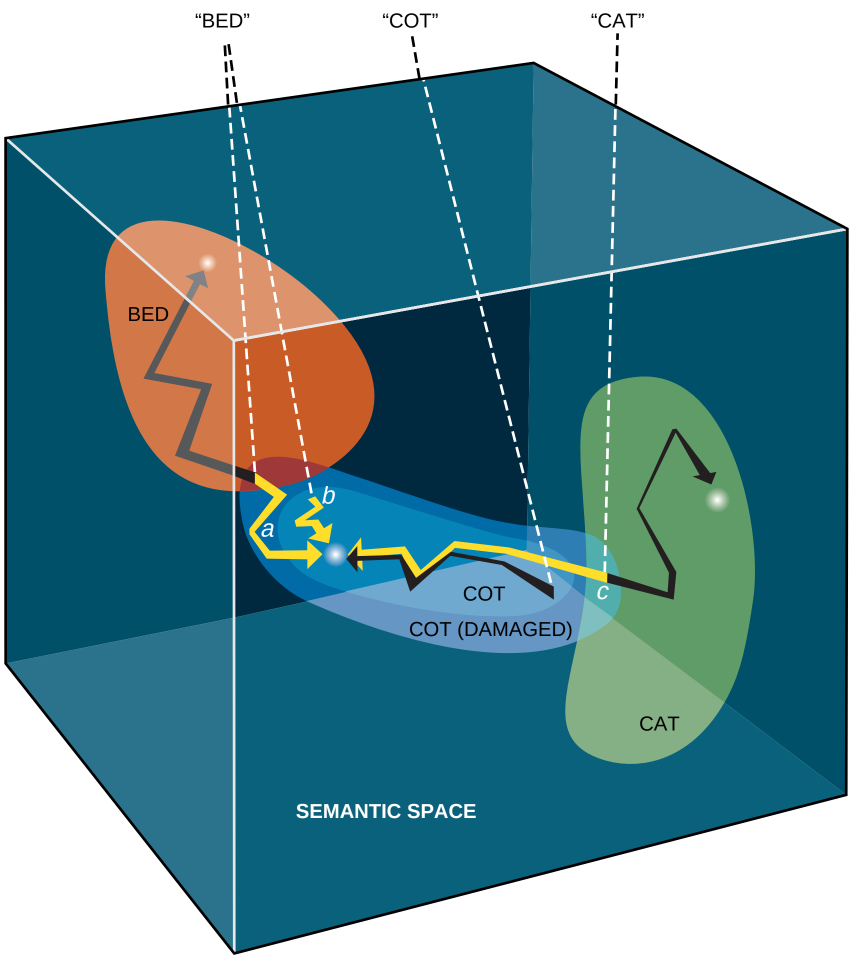
Most of the architectural details are irrelevant. The specific way the visual input is represented is not important as long as words that resemble each other visually produce similar patterns of activity in the input layer. The only crucial ingredient is the existence of attractors—if there are no cleanup units "downstream" of the damage, the network does not exhibit the pattern of errors characteristic of deep dyslexia.

Interestingly enough, our network not only reproduces the obvious visual and semantic errors of deep dyslexia, it also mimics some of the subtler characteristics of the disorder. For instance, patients occasionally make "visual then semantic" errors, in which a semantic confusion seems to follow a visual one. G.R. would read *sympathy* as "orchestra" (presumably via *symphony*). Our networks also produce these errors—sometimes reading *cat* as "bed," via *cot*.

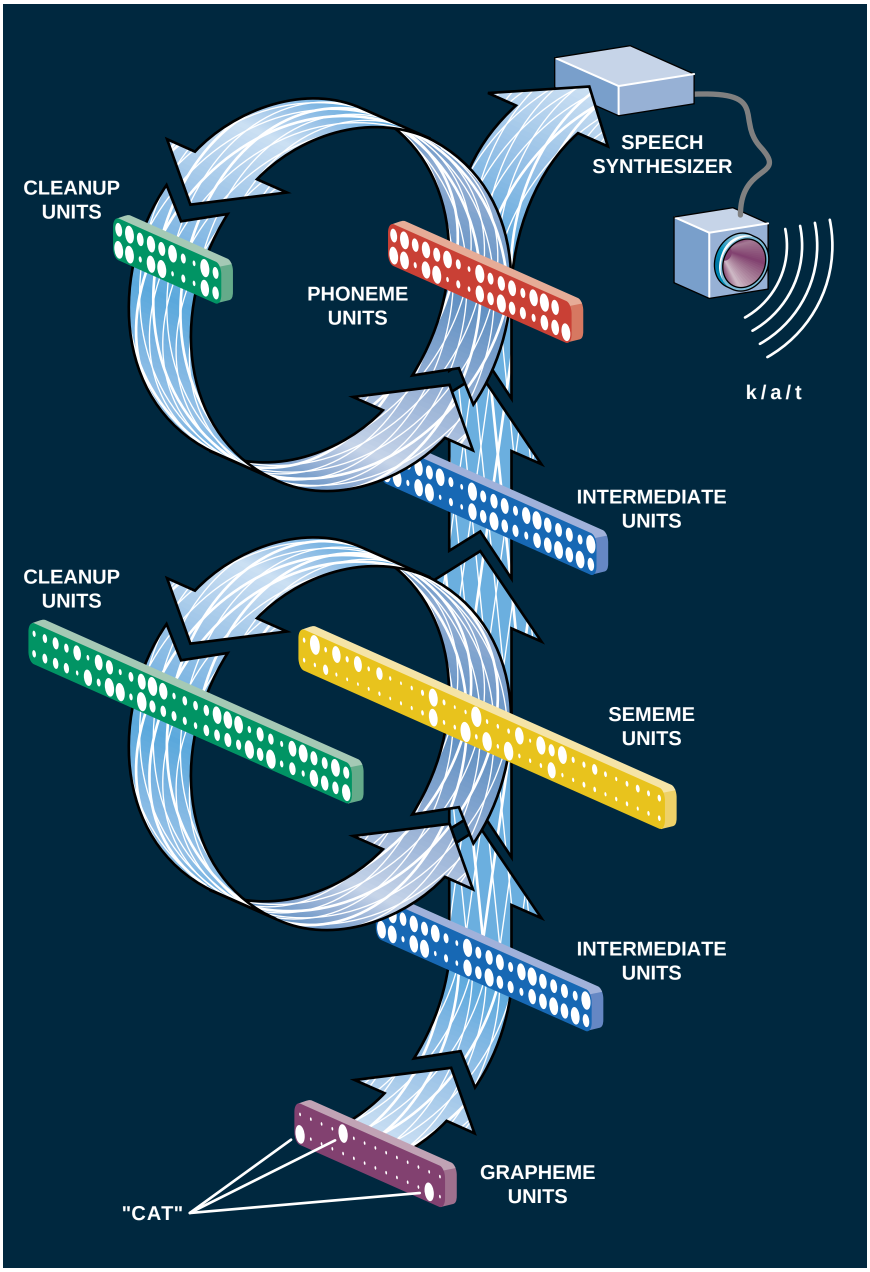
When severely damaged, our network also exhibits a strange effect that occurs when patients have a lesion so large that their semantic representations are distorted beyond recognition and they cannot find a word at all. Such patients are unable to identify the word they are trying to read, but they can often still decide which category it falls into, say, "animal" versus "food." Under similar circumstances, our network no longer stabilizes at the attractor corresponding to a particular word—indeed, the attractors for several words may have merged. Nevertheless, the network's output does stabilize within a larger volume of semantic space wherein the correct word and its relatives once resided. Consequently, the word's category can still be determined.

One symptom of deep dyslexia that our models did not initially address is the way in which patients have more trouble reading abstract words than concrete ones. This phenomenon appears to be an integral part of the syndrome because abstractness—a semantic property—increases the probability of visual errors. Furthermore, when patients make such misreadings, the responses they come up with tend to be more concrete than the original word presented.

We based our approach to simulating this effect on the proposal, made by Gregory V. Jones of the University of Warwick in England and others, that concrete words are easier for deep dyslexic patients because they evoke a more consistent and detailed meaning. In terms of our network, a concrete word has more semantic features than does an abstract one. For example, *post* has 16 features ranging from "size between one foot and two yards" to "used for games or recreation." In contrast, *past* has only two features: "has duration" and "refers to a previous time." We




**SEMANTIC SPACE has many dimensions, corresponding to the semantic features (only a three-dimensional approximation is drawn here). The meanings of particular words are points in semantic space. When the authors' neural network reads a word, interaction between sememe and cleanup units causes any word-form that is mapped into a region of semantic space near the meaning of a word (*colored regions*) to converge on that meaning (*dots*). If the network is damaged so that the boundaries of these so-called attractors shift, a word can be misread as a semantically similar one—"cot" for *bed*, for example (*a*). Semantic errors may also occur if damage causes a word-form to be mapped to a slightly different point in semantic space (*b*). Such a network can make visual errors because visually similar words will initially be mapped to nearby points in semantic space, even if the stable points of the attractors they fall into are quite distant (*c*).**




**SPEAKING NETWORK adds another set of three layers to the original reading neural net. It converts sequences of letter shapes to semantic representations and maps those in turn to sequences of phonemes that can be fed to a speech synthesizer. This network is particularly useful because it does not require researchers to make potentially biased judgments about what word (if any) corresponds to a perturbed pattern of semantic features, as may be generated when the network is damaged to simulate dyslexia.**

designed a new vocabulary containing 20 pairs of four-letter words differing by a single letter, one concrete and the other abstract. On average, the concrete words had about four times as many semantic features as did the abstract ones.

After the network had been trained to pronounce the words, we found that lesions to any part of the network "upstream" of the cleanup units reproduced the effects of abstractness. The concrete words cause fewer errors because there is more redundancy in their semantic activity patterns. Hence, there is more structure that the cleanup units can use to make the network converge on the proper meaning. The abstract words, which have less redundancy in their semantic patterns, must rely more heavily on the feed-forward pathway, where visual influences are the strongest.

Because correct recognition of concrete words relies more on the cleanup circuit, severe damage there leads to a surprising reversal: the damaged network reads concrete words less well and produces more visual errors than with abstract words. This type of lesion and pattern of performance are consistent with what is known about the single, enigmatic patient with "concrete-word dyslexia," studied by Elizabeth K. Warrington at National Hospital in London. Not only did he have much more trouble reading concrete words than abstract ones, he also did better matching spoken abstract words with pictures. This consistency suggests that his problem lay at the level of the semantic system.

Our account of the error pattern of deep dyslexia relies on the properties of a neural network that transforms one representation (a visual word-form) into another, arbitrarily related representation (a set of semantic features). One would expect similar error patterns to result from damage to other cognitive processes that involve an arbitrary transformation to or from a semantic space. Moreover, neuropsychologists have already described somewhat similar error patterns in deep dysgraphia, a disorder of writing, and deep dysphasia, a disorder of word repetition.

This additional evidence suggests that our model may have a wider validity than we originally supposed. More important, however, it marks the successful use of a new technique for understanding how the brain works. Our work differs from other explanations for deep dyslexia (and, with few exceptions, other explanations for neuropsychological phenomena in general) in the kinds of hypotheses that we frame. Instead of verbally characterizing each component in a complex neural mechanism and relying on intuition to tell us how damage will affect its behavior, we simulate that mechanism, damage it and watch to see what happens. We have found that many of our hunches were wrong. This discovery suggests that detailed computer simulations will play a crucial role in furthering understanding of how the brain normally processes information about language and of how that function is disrupted by injury or disease.

FURTHER READING

LESIONING AN ATTRACTOR NETWORK: INVESTIGATIONS OF ACQUIRED DYSLEXIA. G. E. Hinton and T. Shallice in *Psychological Review,* Vol. 98, No. 1, pages 74–95; January 1991.

DEEP DYSLEXIA: A CASE STUDY OF CONNECTIONIST NEUROPSYCHOLOGY. D. C. Plaut and T. Shallice in *Cognitive Neuropsychology,* Vol. 10, No. 5, October 1993.

# Raising the *Vasa*

*This Swedish man-of-war foundered on her maiden voyage and slept for three centuries at the bottom of Stockholm Harbor. Here is the story of her resurrection*

by Lars-Åke Kvarning

In the 1620s Gustavus II Adolphus, king of Sweden, was on his way to becoming one of the leading warlords in Europe. He controlled Finland, Estonia and Livonia, and he had just won the small part of Russia that touches on the Gulf of Finland. By thus excluding the czar from the Baltic, he had nearly made that sea into a Swedish lake. To secure his power, he needed to maintain a strong fleet in the Baltic. Therefore, in January 1625, the king commissioned the construction of four ships. The most magnificent was the *Vasa*. Named after the Swedish royal family, she was one of the largest warships of the time.

On August 10, 1628, a bright, sunny day, crowds gathered at Stockholm Harbor to see this new addition to the fleet leave for the island naval base not far from the city. Members of the crew went aloft to set four of the *Vasa*'s 10 sails—foresail, foretop, maintop and mizzen. Cannons fired salutes, and the cheering throng on the quays and shores watched as the imposing vessel began to move slowly in the light breeze, her colors and gilding shining.

Suddenly the cheers changed to shrieks of horror as a gust of wind caught the ship in the harbor. She heeled over to port and raised a little only to list again, the water now rushing in through the open gunports. Then the *Vasa* foundered, "under sails, pennants and all," as the Council of the Realm described the disaster in a letter to the king. The Danish ambassador, Erik Krabbe, reported that the drowned "are said to be more than half a hundred persons, among them some women and children who wanted to follow their husbands into the archipelago."

Three centuries later the wreck of the *Vasa* was found, raised and restored, making the ship a monument to her builders and rescuers alike. The warship's resurrection was worthwhile both for the intrinsic interest of the operation itself and because the prize has so much to tell about the vessel's makers. A ship is not only a society in miniature; it is also the product of a larger society.

The loss of one of the biggest warships in Europe, on her maiden voyage and in sheltered waters, constituted a catastrophe—one that called for a scapegoat. Interrogations of the captain of the ship, Söfring Hansson, and other officers began soon after they were rescued. The shipbuilder, Henrik Hybertson, had died the year before, and the responsibility of completing the project had been taken over by his widow and brother in cooperation with the assistant shipbuilder, Hein Jacobson. These people also had to submit to questioning.

Had the guns been properly secured? The principals could not prove it, but in the summer of 1961 marine archaeologists indeed found the cannon carriages standing in their proper places, the remains of their securing ropes still coiled around the axles. Had the stone ballast been taken on properly? A witness replied in the affirmative that it had been and that the space for it down in the hold was filled.

Signs of trouble may be read between the lines, however. The skipper, Göran Mattson, related that the vice admiral and the captain of the *Vasa* had tested the stability of the ship while she was being fitted, by having 30 men run to and fro across the upper deck. After the third cycle, they had had to call a halt to prevent the *Vasa* from capsizing on the spot. Something was clearly wrong with the vessel's design, but still no action was taken.

It can be gathered from the records that most observers believed the ballast compartment had not been made large enough to balance the extraordinary double row of gundecks that the king had specified. Indeed, the shipwrights maintained that they had built

LARS-ÅKE KVARNING is president of Sweden's National Maritime Museums and director of the Vasa Museum in Stockholm, where he has worked since 1964. Before that he was associated with the folk department of the Nordic Museum and with the Swedish Central Office of National Antiquities. He holds a licentiate in ethnology from Uppsala University and has directed a documentary project to photograph national monuments from the air. He serves as a reserve captain in the life guard regiment of the Grenadiers.

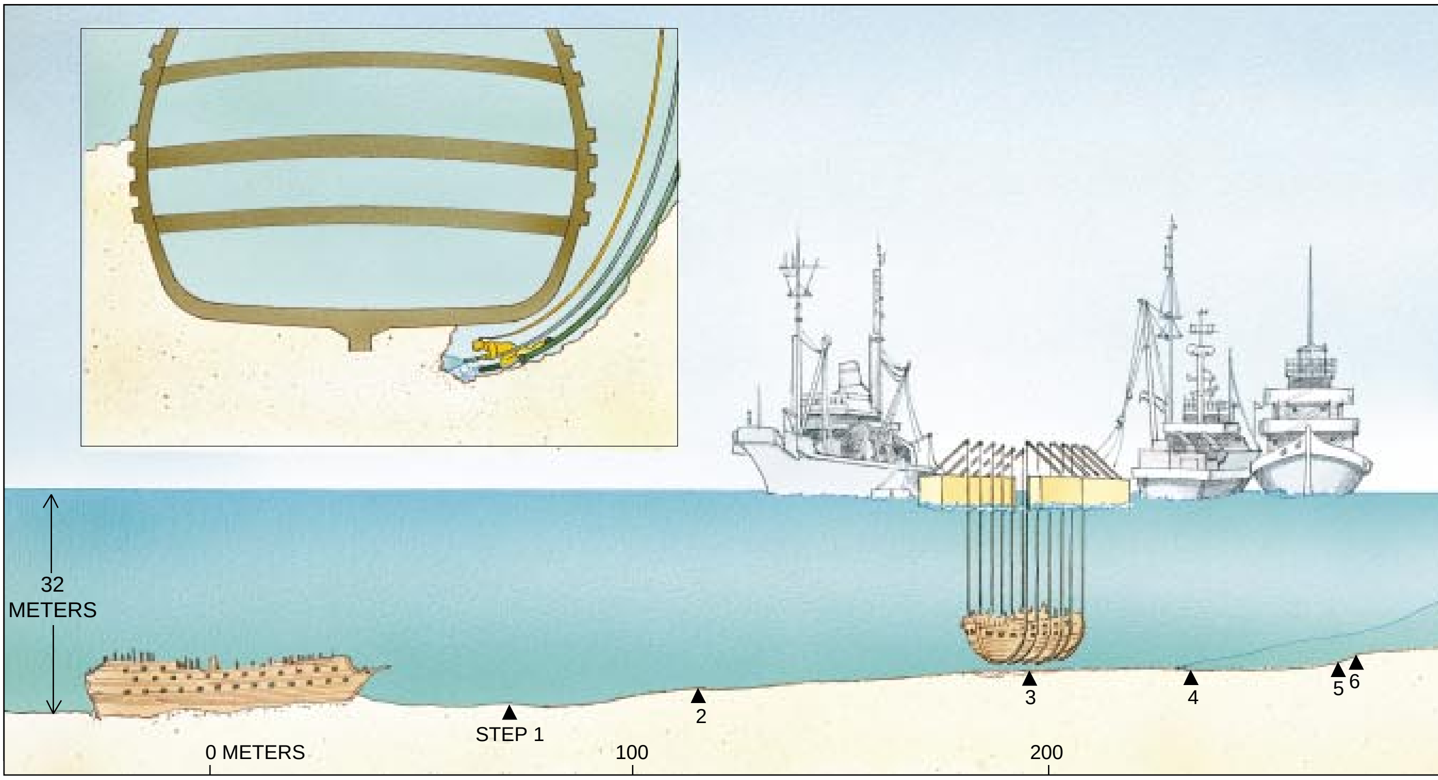
**RESTORED MAGNIFICENCE of the *Vasa* is apparent in the stern, which sports such royal symbols as a crown, a coat of arms and griffins.**

the vessel just as contracted and in accordance with orders received directly from the king.

Gustavus Adolphus had wanted a strong navy to protect the supply lines leading to the theater of war on the other side of the Baltic. He also intended to blockade the rich Polish ports so that his agents could collect customs duties from cargo ships calling there. It seems that his ambitions grew after he placed his original order and that the *Vasa* was begun as a smaller ship and completed as a larger one.

This royal intervention was perhaps the real reason for the disaster. The




shipwrights of the day did not make construction drawings and had no means of calculating the stability of their creations. Trial and error was their only guide. When the builders scaled up the plans of the *Vasa* by adding a second gundeck, they left only enough room in the hold for about 121 tons of stone ballast, less than half of what was needed. Yet had they attempted to increase the amount, the lower gundecks would have come down dangerously close to the waterline.

Many others made mention of the king's personal interest in the building of the vessel. The shipwright, when asked about whom or what to blame for the debacle, sighed: "Only God knows...." With God and the king involved in this fashion, there was only one way out. The case was dropped.

A ship of this size, equipped with bronze cannons worth a fortune, could not but become an immediate object of salvage. Only three days after the calamity, the Council of the Realm authorized Ian Bulmer, an Englishman, to make the attempt. He achieved little. Although the principles of such operations were well known, the salvage crews lacked the force to lift a prize as heavy as the *Vasa*.

A Swede named Albrect von Treileben came on the scene in 1663, five years after introducing a diving bell that made submarine sites far more accessible. Thus equipped, he and his divers set out to salvage the cannons and other loose items of value. A diver would don leather clothing to protect against the cold and climb on a platform hanging under the bell, his head and chest projecting into the bubble of air. Then the bell would be lowered down to the sunken wreck.

Francesco Negri, an Italian priest and explorer, has left posterity this eyewitness account of von Treileben's operations:

> When he had signaled and been hauled up after having been under water at least a quarter of an hour, he brought up with him a heavy plank of oak with iron mountings that he had caught with his hook. I asked him through an interpreter if he could have stayed down any longer. He said yes, up to half an hour, but no longer. It was possible—as I believe, although I forgot to ask him—that the air enclosed by the bell would have become too hot from his breathing, and probably the water would have excessively tormented his legs with cold. The man trembled, although he was born in the country and was strong and used to physical strain. It was then towards the end of October of last year, 1663.

When below, the diver could hardly move and could not see at all. Still, the salvagers managed to bring up most of the *Vasa*'s 64 cannons, including 48 24-pounders weighing more than a ton each. The remainder of the wreck had little to interest the people of the day, and so they abandoned it to the sea, ending a salvage operation that was perhaps the most impressive underwater enterprise of the premodern era.

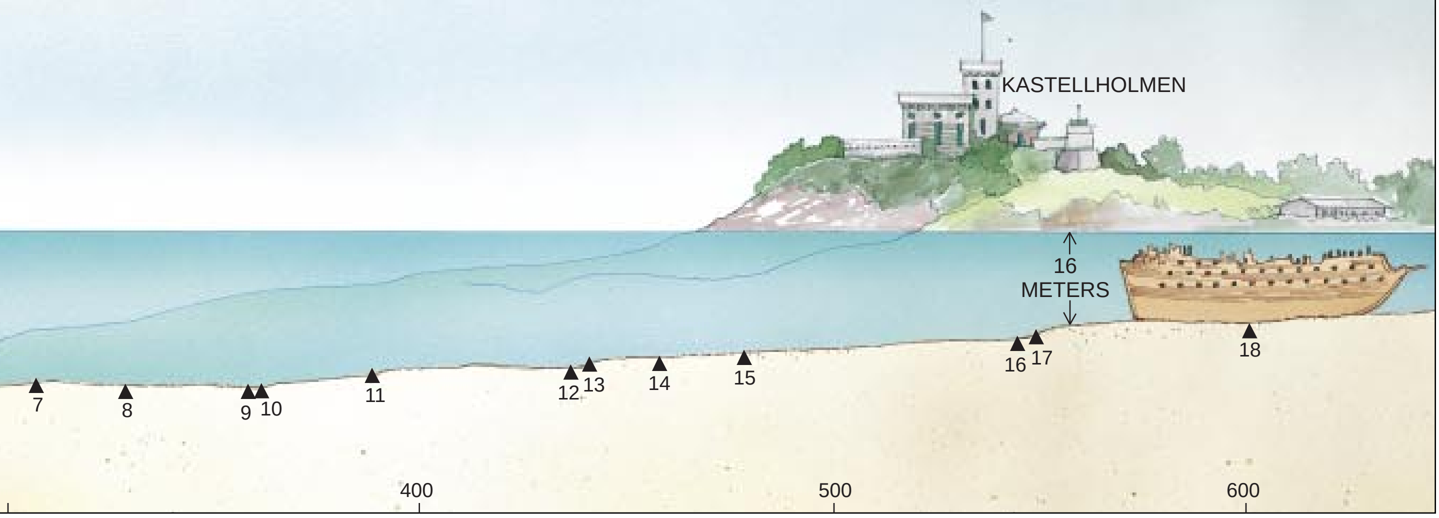
The divers were lucky as well as courageous. The limited supply of air cut the dives short enough to prevent dangerous quantities of nitrogen from dissolving, under high pressure, in the men's blood. A too rapid return to the surface would then have caused the gas to precipitate in bubbles, an often fatal malady commonly known as the bends.

The *Vasa* was more or less forgotten until 1956, when Anders Franzén, a private researcher, located her on the bottom of Stockholm Harbor. Franzén owed his success not to chance but to insight. He realized that the Baltic was not salty enough to support the shipworm *Teredo navalis*—an elongated, wormlike clam—and that its waters therefore ought to have preserved the wrecks of wooden ships extraordinarily well. The Baltic's spoils are indeed without parallel, for no other sea of such low salinity has borne so many fleets.

Franzén concentrated on men-of-war from the 16th and 17th centuries. He studied the archives and got indications of the positions of 12 sunken ships that he found particularly interesting and then set his mind on finding the *Vasa*. One day, while dragging his grappling iron in Stockholm Harbor, it hooked something big and immovable. Divers went down and confirmed that an enormous hull of black oak was

## Salvaging the *Vasa*

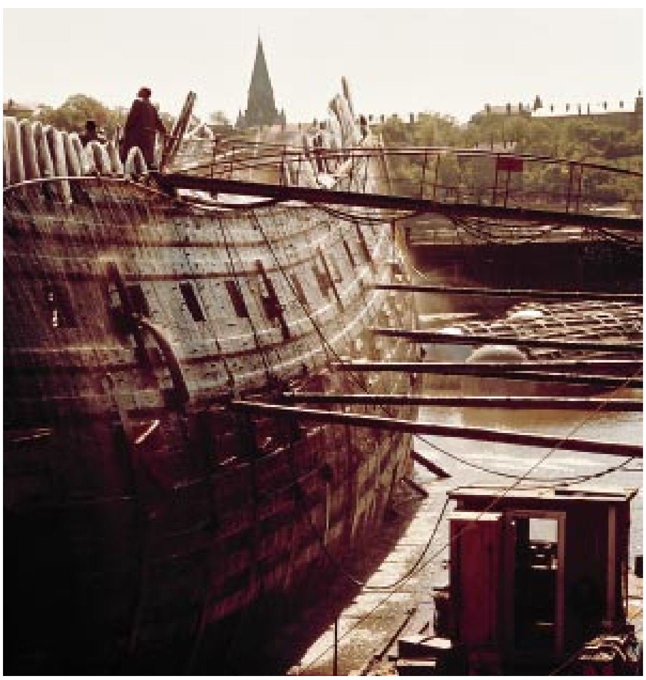
Divers armed with water jets blasted six tunnels through the clay beneath the sunken wreck (*inset at left*). They then drew steel cables through the tunnels, fixed either end of each cable to parallel banks of half-submerged pontoons and tightened them, forming a taut cat's cradle. By bailing out the pontoons, the workers could lift the wreck 2.5 meters—about eight feet—so it could be towed shoreward. When the ship set down again, the workers refilled the pontoons, tightened the cables and repeated the process. After 18 steps, they reached shallower water. There, after two years of preparation, jacks mounted on the pontoons finally lifted the ship to the surface.




standing on its keel on the bottom. It was the *Vasa*. In and around her were strewn the skeletons of 25 people. A few of them were found inside the ship during the excavation, but most were turned up by workers who scoured the underlying seafloor after the vessel had been raised.

Franzén's dream of recovering the ship invited ridicule from several quarters, but he was able to attract two weighty collaborators. The Swedish Navy agreed to organize the training of its divers around the needs of the project, and the Broströms Shipping Company, through its subsidiary the Neptun Salvaging Company, undertook the task free of charge.

There was no lack of imaginative suggestions. One wag proposed using Ping-Pong balls to float the *Vasa* to the surface. Another recommended freezing the ship in a block of ice, which would then bob to the surface, where it could

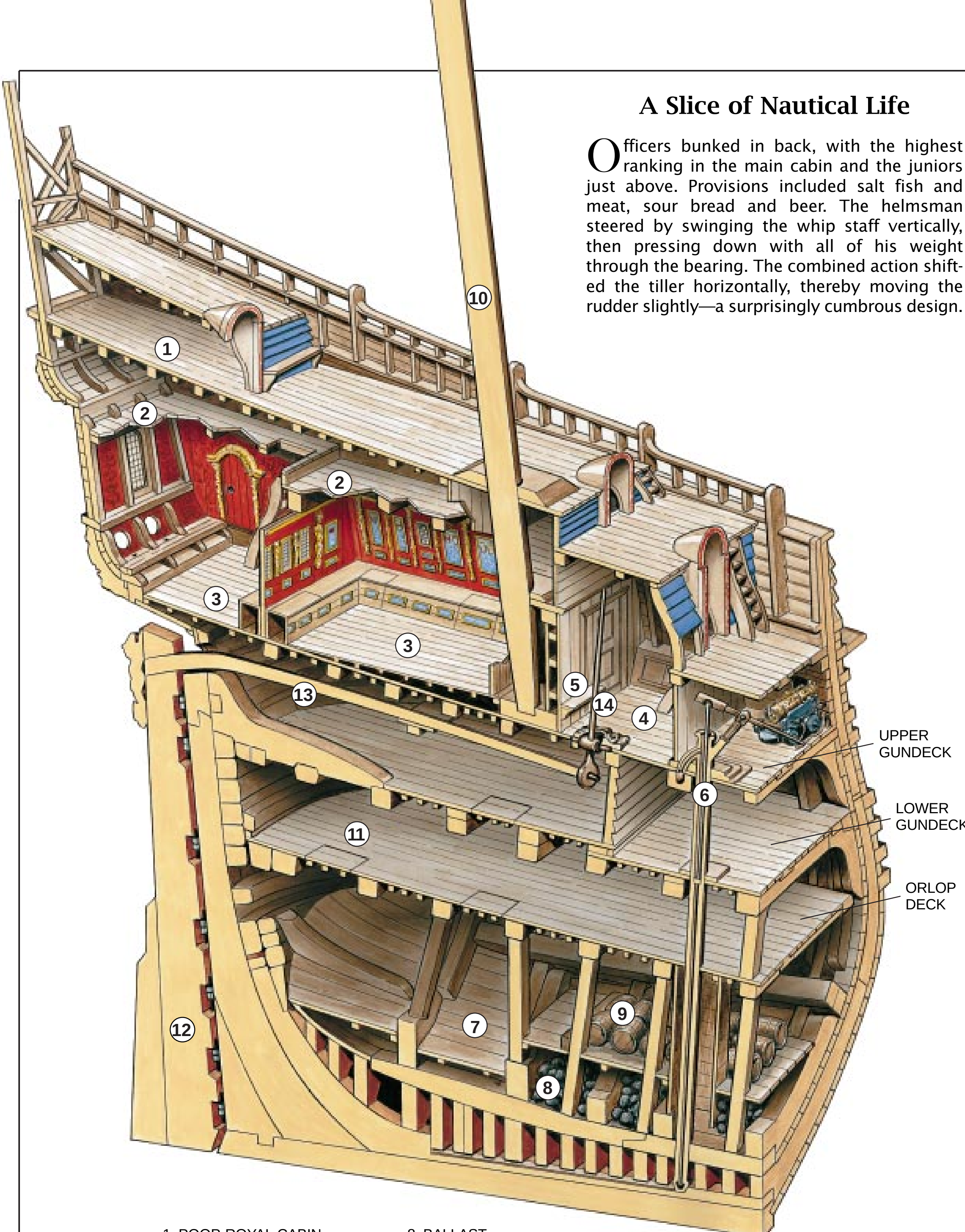


**RAISED AT LAST, the *Vasa*'s hull was still held together by the original wooden pegs, although all the iron bolts and nails had rusted away many years before.**

**SHOWER OF POLYETHYLENE GLYCOL impregnated the *Vasa* over a period of 18 years. Smaller objects were treated more quickly by immersion in baths.**

# A Slice of Nautical Life

Officers bunked in back, with the highest ranking in the main cabin and the juniors just above. Provisions included salt fish and meat, sour bread and beer. The helmsman steered by swinging the whip staff vertically, then pressing down with all of his weight through the bearing. The combined action shifted the tiller horizontally, thereby moving the rudder slightly—a surprisingly cumbrous design.




1. POOP-ROYAL CABIN
2. OFFICERS' CABIN
3. MAIN CABIN
4. STEERAGE CABIN
5. HELMSMAN'S STATION
6. BILGE PUMP
7. POWDER MAGAZINE
8. BALLAST
9. PROVISIONS STORE
10. MIZZEN MAST
11. BARBER-SURGEON STATION
12. RUDDER
13. TILLER
14. WHIP STAFF

be towed to a suitable place for melting.

The salvage company, however, favored a traditional technique. It ran cables under the *Vasa,* sparing her timbers by snaking the cables through tunnels bored in the seabed by two-faced water guns, known as *Zetterström* nozzles. The forward jet cuts through the clay while weaker backward jets offset the recoil and clear away loosened material. Divers had to work within the growing tunnels, with four meters of mud and sludge, a 333-year-old ship and 32 meters of water directly overhead. Their worst nightmare, as they toiled in the complete darkness, was that the ship's bottom would give way under the weight of the ballast stones inside, trapping them in their holes.

It was a breathless moment when the *Vasa* at last sat within her steel cat's cradle, ready for the tug that would loosen her from the clay that had imprisoned her for so long. Pumps emptied pontoons attached to the cables, which floated up, tightening the cables until they were as tense as violin strings. The clay at last let go its grip, and the *Vasa* hung free in the cables. This suspension put a tremendous strain on the hull, and it would have collapsed if not for the strength remaining in its ancient oaken timbers. Now everything went very smoothly.

Because the pontoons could not lift the *Vasa* more than 2.5 meters at a time, the process had to ratchet through 18 steps. The *Vasa* would rise a bit, then the workers would tow her to shallower water and flood her pontoons until she set down on the bottom again. Remarkably, the impressions the *Vasa* left on the seafloor at each stage are clearly visible today, appearing like a giant's footsteps on the sonar scan.

At last the ship reached a site from which the final lift could be made with ease. Huge jacks mounted on the pontoons hauled the prize to the surface, at which point pumps that had been installed inside the hull began to bail it out. Thus assisted, the *Vasa* floated on her own keel into dry dock, where a concrete pontoon was waiting. There, in 1961, for the first time in 333 years, the *Vasa* revealed her imposing hull to an astonished public.

To protect the delicate, waterlogged wood, the caretakers of the man-of-war constantly sprayed her with water until permanent preservation could be arranged inside an appropriate shelter. It was also necessary to get rid of the vast amount of mud that had collected in the ship—a black sludge loaded with archaeological finds and, perhaps, bacteria. To be on the safe side, the archaeologists who were to excavate the vessel from within were vaccinated against tetanus, typhus and other communicable diseases. Thus protected, they found and registered about 25,000 objects, about half of them deriving from the ship's structure.

All the finds were to return to their proper places during reconstruction, the aim of which was to restore the *Vasa* as far as possible to her original shape. To that end, in the fall of 1961 an aluminum shelter was constructed around the *Vasa,* which still rested on the pontoon. The pontoon was then towed to a temporary museum, where work could proceed under the eyes of the public. The challenge was to replace the water in the wood with something more lasting without allowing the wood to crack, shrink, warp or misbehave in any way. Not only the 1,100 tons of structural components had to be thus preserved but also 700 sculptures and carved ornaments and many thousands of objects made of textiles, leather and metal.

No one had ever attempted to preserve such a mass of wood. But two Swedes—Bertil Centervall, a conservator, and Rolf Morén, an engineer—had invented a process to preserve fresh wood by impregnating it with polyethylene glycol (PEG). Because the polymer worked by replacing water in fresh wood, there was reason to hope that it might do the same in ancient waterlogged wood—and it did. Workers dissolved the PEG in water, added borax and boric acid as a safeguard against fungus and applied the mixture to the wood. The PEG effectively embalms the wood by stiffening its cells, thus preventing the wood from changing its shape or cracking. Additional protection was provided by climate-control systems that assured a constant temperature and an ideal humidity.

Wooden objects of manageable size went into big baths, where the concentration and temperature of the solution could be increased slowly. This very efficient form of the treatment took 18 months for the oak and about a year for the softwoods. The ship, however, could not be dipped in this fashion. The workers had to sprinkle the solution on her, inside and outside, a tedious process of saturation that started in 1961, three years before I joined the restoration team, and ended only in 1979. At that point, the hull was slowly dried.

An intriguing preservation problem was offered by the six sails that survived from the *Vasa's* original 10. They were discovered in a sail locker on the orlop, the lowest continuous deck. The cloth was in such bad condition that mere handling might have caused it to crumble into dust. It was possible, however, to work the material as long as it was kept in water. Large baths were therefore prepared in which the remains of the sails could be unfolded, cleaned and dried in alcohol and xylene. Because of their fragility,

**MODEL OF THE *VASA,* constructed at one tenth the original size, took four craftsmen four years to complete. It now sits alongside the reconstructed hull in the Vasa Museum, which faces Stockholm Harbor.**





Copyright 1993 Scientific American, Inc.

**SMALL OBJECTS afford perhaps the clearest view of life on board the *Vasa.* The vessel's brilliant colors are evidenced in the fully restored heraldic device, whose sheaf represents the word *Vasa,* the name of the Swedish ruling family (*far left, top*). Other items include (*clockwise from left*) clothing, mess kits, officers' tableware, apothecary materials, sailmaking equipment, ammunition, games, coins and toiletries.**

they were affixed to a backing of glass fibers by means of a plastic solution made especially for this purpose. The solution has the same refractive index as the glass fibers, making them invisible.

The wrought iron had long since rusted away, with the exception of the lowest rudder pintle and gudgeon (the pin and the pivot on which the rudder had turned), which had been pushed down into the bottom clay and preserved in relatively good condition. In 1992 we also found a hidden bolt in very good condition when we were preparing for the masts to be restepped.

The cast-iron cannonballs were protected by their low carbon content and thus resisted oxidation more effectively. Many of them now weighed hardly more than a tennis ball, but their rusted hulks retained their original shape and size. Because these rust balls would have fallen apart in air, we dried them in hydrogen at 1,060 degrees Celsius, thereby restoring the iron while purging injurious compounds.

The next problem consisted in putting the ship back together. Fortunately, the hull was built with a double system of iron bolts and wooden pegs, so that the disappearance of the bolts did not cause the wooden planks to collapse in a heap. The main part of the hull could therefore come up in one piece. The stern castle, however, had broken down, and the beak head—a protruding construction at the prow—and other parts had fallen off. Consequently, workers had to identify and locate many thousands of structural components, ranging from heavy beams to tiny bits of wood—a gigantic jigsaw puzzle to be assembled without benefit of blueprints.

The restorers hesitated over the propriety of filling in gaps in the archaeological record. At first they were able to dodge the question because such a wealth of original parts remained. But soon they reached a stage at which connecting structures had to be made to exploit some of the surviving pieces. By then, however, we realized that we wanted to rebuild the ship as completely as possible and that very few missing parts would have to be supplied by our own people. Symmetry helped enormously: a ship is the same on either side, so when anything was missing we were usually able to model it on its counterpart from the other side. Of course, we always mark such reconstructions.

The restored *Vasa* has revealed many unexpected aspects of the shipwright's craft. The steering system, for instance, is strangely awkward. The tiller stretched 11 meters from the top of the rudder to the eye of the whip staff—a vertical beam that connected at a right angle to the tiller and projected up to the deck through a roller bearing. To steer, the helmsman first had to tilt the whip staff and then bear down, pressing with all his weight through the bearing. The combined action would move the tiller to the side. To center the rudder, he had to haul the whip staff back up through the hole in the bearing.

It appears that the shipwright was following a tradition that had made sense only when vessels were smaller. In the *Vasa,* the great length of the tiller limited the turning radius of the rudder to 14 degrees—seven to the left and seven to the right. That paltry distance, combined with the relatively small surface of the rudder, meant that most of the steering had to be done by manipulating the sails.

The most striking feature of the *Vasa* is her gorgeous ornamentation, so different from the modern subordination of form to function. The *Vasa* may well illustrate the change in taste more effectively than can architectural remains, for one expects to see decoration in buildings but not in military hardware.

Yet the *Vasa* is festooned throughout with sculptures prominent enough to attract the eye during an engagement. They filled several purposes: to encourage friends, intimidate enemies, assert claims and impress the world with this picture of power and glory.

The decoration of the stern, for instance, spun a story of the Vasa, the Swedish royal family for whom the ship was named. The ornamentation begins at the top with a woodcut of the young King Gustavus Adolphus, a royal crown held over his head by two griffins. This is a political statement, for it contradicts King Sigismund III of Poland—Gustavus Adolphus's first cousin—who also laid claim to the Swedish crown. The statement is reinforced by the frieze just below, which reads "G A R S," for Gustavus Adolphus Rex Sueciae, designating Gustavus Adolphus king of Sweden. Further down, the Swedish national coat of arms appears and, below it, the family crest of the Vasa dynasty. *Vasa* is an old Swedish word for "sheaf," pictured on the coat-of-arms.

Hardly any traces of color remained on the waterlogged sculptures, the wood of which had turned to a rather even dark brown. On closer inspection, bits of gold leaf still glittered here and there, leading us to believe that many of the sculptures had been gilded. We were doubtful as to other colors, but a number of applicable techniques to discern them have been developed since 1961, notably the combination of an electron microscope and an energy-dispersive x-ray analyzer. The *Vasa,* it is clear, was as bright with colors and as resplendent with gold as the altarpiece of a baroque church. So far we have used this information to restore only a few sculptures, among them the Vasa family crest.

Life on board the warship would have been hard, even in the 1620s, when few sea battles were waged. But peace also had its horrors: more than 400 men would have been crammed in a confined space. Hygienic conditions on the ships of the day were shocking, and the food was bad even in port, let alone after stewing in the hold for a few weeks. The men and the officers, too, were thus exposed to a range of diseases.

A sunken ship is a kind of time capsule. Indeed, it is to be preferred to the contrivances of the cornerstone cache and burial chamber because it provides an utterly unself-conscious snapshot of daily life. Ships, their equipment and their cargoes contain a wealth of information about technology and commerce, wars and diplomacy, common people and the spirit of the age.

The *Vasa* was not fully manned and equipped when she went down. The ship lacked, for instance, 300 marines from the normal military complement. Nevertheless, thousands of personal belongings were recovered. The simple sailors had things of wood or clay for the most part. The officers' quarters in the stern castle contained delicate Swedish glassware, porcelain, pewter and brass, as well as goods imported from other parts of Europe. One pewter flask still contained clear alcohol, 66 proof, that seems to be arrack from the East Indies. I can testify, from personal experience, that the liquor was good.

In June 1960, King Carl XVI Gustaf inaugurated the Vasa Museum in Stockholm. There the royal warship rests, surrounded by exhibits of the world from which she comes—a graphic testimonial to another time and place.


FURTHER READING

THE SWEDISH WARSHIP WASA. Lars-Åke Kvarning and Bengt Ohrelius. Macmillan, 1974.

WHY WASA CAPSIZED. Curt Borgenstam and Anders Sandström. Wasa Studies No. 13. National Maritime Museum, 1984.

THE POWER AND THE GLORY: THE SCULPTURES OF THE WARSHIP WASA. Hans Soop. Almqvist and Wiksell/Coronet Books, 1986.

THE ROYAL WARSHIP VASA. Björn Landström. Stenström Interpublishing, 1988.

VASA. Erling Matz. Vasa Museum, 1990.

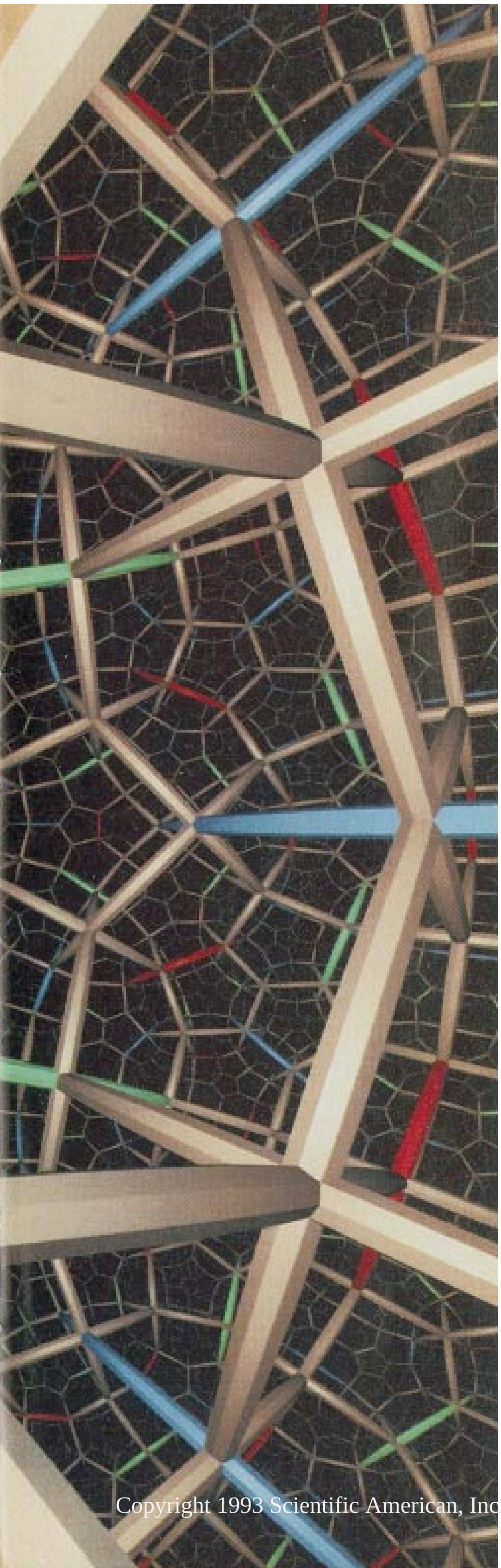
THE VASA MUSEUM: AN OLD SHIP IN A NEW HOUSE. K. A. Adrup. National Board of Public Building, 1990.

TRENDS IN MATHEMATICS

# THE DEATH OF PROOF

by John Horgan, *senior writer*

**Computers are transforming the way mathematicians discover, prove and communicate ideas, but is there a place for absolute certainty in this brave new world?**



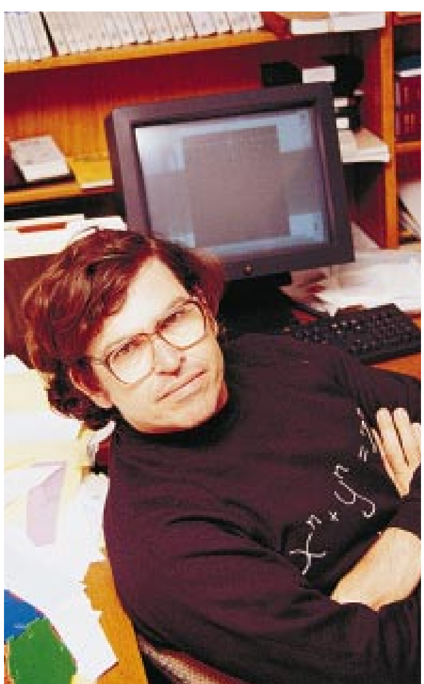


Legend has it that when Pythagoras and his followers discovered the theorem that bears his name in the sixth century B.C., they slaughtered an ox and feasted in celebration. And well they might. The relation they found between the sides of a right triangle held true not sometimes or most of the time but always—regardless of whether the triangle was a piece of silk or a plot of land or marks on papyrus. It seemed like magic, a gift from the gods. No wonder so many thinkers, from Plato to Kant, came to believe that mathematics offers the purest truths humans are permitted to know.

That faith seemed reaffirmed this past June when Andrew J. Wiles of Princeton University revealed during a meeting at the University of Cambridge that he had solved Fermat's last theorem. This problem, one of the most famous in mathematics, was posed more than 350 years ago, and its roots extend back to Pythagoras himself. Since no oxen were available, Wiles's listeners showed their appreciation by clapping their hands.

But was the proof of Fermat's last theorem the last gasp of a dying culture? Mathematics, that most tradition-bound of intellectual enterprises, is undergoing profound changes. For millennia, mathematicians have measured progress in terms of what they can demonstrate through proofs—that is, a series of logical steps leading from a set of axioms to an irrefutable conclusion. Now the doubts riddling modern human thought have finally infected mathematics. Mathematicians may at last be forced to accept what many scientists and philosophers already have admitted: their assertions are, at best, only provisionally true, true until proved false.

This uncertainty stems, in part, from the growing complexity of mathematics. Proofs are often so long and complicated that they are difficult to evaluate. Wiles's demonstration runs to 200 pages—and experts estimate it could be five times longer if he spelled out all its elements. One observer as-



*"VIDEO PROOF" dramatizes a theorem, proved by William P. Thurston of the Mathematical Sciences Research Institute* (left), *that establishes a profound connection between topology and geometry. The theorem shows how the space surrounding a complex knot (represented by the lattice in this scene) yields a "hyperbolic" geometry, in which parallel lines diverge and the sides of pentagons form right angles. The computer-generated video, called Not Knot, was produced at the Geometry Center in Minnesota.*

serted that only one tenth of 1 percent of the mathematics community was qualified to evaluate the proof. Wiles's claim was accepted largely on the basis of his reputation and the reputations of those whose work he built on. Mathematicians who had not yet examined the argument in detail nonetheless commented that it "looks beautiful" and "has the ring of truth."

Another catalyst of change is the computer, which is compelling mathematicians to reconsider the very nature of proof and, hence, of truth. In recent years, some proofs have required enormous calculations by computers. No mere human can verify these so-called computer proofs, just other computers. Recently investigators have proposed a computational proof that offers only the probability—not the certainty—of truth, a statement that some mathematicians consider an oxymoron. Still others are generating "video proofs" in the hopes that they will be more persuasive than page on page of formal terminology.

At the same time, some mathematicians are challenging the notion that formal proofs should be the supreme standard of truth. Although no one advocates doing away with proofs altogether, some practitioners think the validity of certain propositions may be better established by comparing them with experiments run on computers or with real-world phenomena. "Within the next 50 years I think the importance of proof in mathematics will diminish," says Keith Devlin of Colby College, who writes a column on computers for *Notices of the American Mathematical Society*. "You will see many more people doing mathematics without necessarily doing proofs."

Powerful institutional forces are promulgating these heresies. For several years, the National Science Foundation has been urging mathematicians to become more involved in computer science and other fields with potential applications. Some leading lights, notably Phillip A. Griffiths, director of the Institute for Advanced Study in Princeton, N.J., and Michael Atiyah, who won a Fields Medal (often called the Nobel Prize of mathematics) in 1966 and now heads Cambridge's Isaac Newton Institute for Mathematical Sciences, have likewise encouraged mathematicians to venture forth from their ivory towers and mingle with the real world. At a time when funds and jobs are scarce, young mathematicians cannot afford to ignore these exhortations.

There are pockets of resistance, of course. Some workers are complaining bitterly about the computerization of their field and the growing emphasis on (oh, dirty word) "applications." One of the most vocal champions of tradition is Steven G. Krantz of Washington University. In speeches and articles, Krantz has urged students to choose mathematics over computer science, which he warns could be a passing fad. Last year, he recalls, a National Science Foundation representative came to his university and announced that the agency could no longer afford to support mathematics that was not "goal-oriented." "We could stand up and say this is wrong," Krantz grumbles, "but mathematicians are spineless slobs, and they don't have a tradition of doing that."

David Mumford of Harvard University, who won a Fields Medal in 1974 for research in pure mathematics and is now studying artificial vision, wrote recently that "despite all the hype, the press, the pressure from funding agencies, et cetera, the pure mathematical community by and large still regards

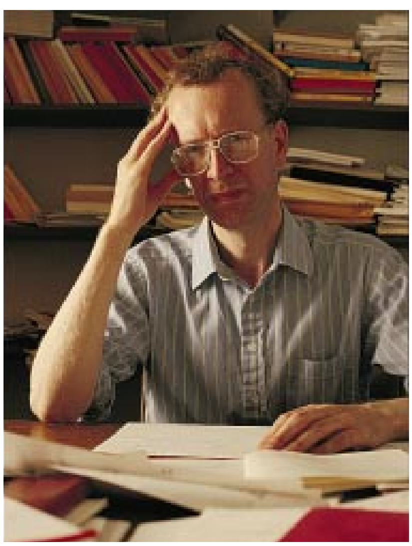
## A Splendid Anachronism?

Those who consider experimental mathematics and computer proofs to be abominations rather than innovations have a special reason to delight in the conquest of Fermat's last theorem by Andrew J. Wiles of Princeton University. Wiles's achievement was a triumph of tradition, running against every current in modern mathematics.

Wiles is a staunch believer in mathematics for its own sake. "I certainly wouldn't want to see mathematics just being a servant to applications, because it's not even in the interests of the applications themselves," he says.

The problem he solved, first posed more than 350 years ago by the French polymath Pierre de Fermat, is a glorious example of a purely mathematical puzzle. Fermat claimed to have found a proof of the following proposition: for the equation $X^N + Y^N = Z^N$, there are no integral solutions for any value of $N$ greater than 2. The efforts of mathematicians to find the proof (which Fermat never did disclose) helped to lay the foundation of modern number theory, the study of whole numbers, which has recently become useful in cryptography. Yet Fermat's last theorem itself "is very unlikely to have any applications," Wiles says.

Although funding agencies have been encouraging mathematicians to collaborate, both with each other and with scientists, Wiles worked in virtual solitude for seven years. He shared his ideas with only a few colleagues toward the end of his quest.



Wiles's proof has essentially the same classical, deductive form that Euclid's geometric theorems did. It does not involve any computation, and it claims to be absolutely—not probably—true. Nor did Wiles employ computers to represent ideas graphically, to perform calculations or even to compose his paper; a secretary typed his hand-written notes.

He concedes that testing conjectures with computers may be helpful. In the 1970s computer tests suggested that a far-fetched proposal called the Taniyama conjecture might be true. The tests spurred work that laid the foundation for Wiles's own proof.

Nevertheless, Wiles doubts he will take the trouble to learn how to perform computer investigations. "It's a separate skill," he explains, "and if you're investing that much time on a separate skill, it's quite likely it's taking you away from your real work on the problem."

He rejects the possibility that there may be a finite number of truths accessible to traditional forms of inquiry. "I disagree vehemently with the idea that good theorems are running out," he says. "I think we've barely scratched the surface."

EXPERIMENTAL MATHEMATICIAN **Jean E. Taylor of Rutgers University seeks the rules governing minimal surfaces by studying real phenomena, such as soap bubbles, and computer-generated ones, such as idealized crystals (*left*).**

computers as invaders, despoilers of the sacred ground." Last year Mumford proposed a course in which instructors would show students how to program a computer to find solutions in advanced calculus. "I was vetoed," he recalled, "and not on the grounds—which I expected—that the students would complain, but because half of my fellow teachers couldn't program!"

That situation is changing fast, if the University of Minnesota's Geometry Center is any indication. Founded two years ago, the Geometry Center occupies the fifth floor of a gleaming, steel and glass polyhedron in Minneapolis. It receives $2 million a year from the National Science Foundation, the Department of Energy and the university. The center's permanent faculty members, most of whom hold positions elsewhere, include some of the most prominent mathematicians in the world.

On a recent day there, several young staff members are editing a video demonstrating how a sphere can be mashed, twisted, yanked and finally turned inside out. In a conference room, three computer scientists from major universities are telling a score of high school teachers how to create computer graphics programs to teach mathematics. Other researchers sit at charcoal-colored NeXT terminals, pondering luridly hued pictures of four-dimensional "hypercubes," whirlpooling fractals and lattices that plunge toward infinity. No paper or pencils are in sight.

At one terminal is David Ben-Zvi, a Harpo Marx–haired junior at Princeton who is spending six months here exploring nonlinear dynamics. He dismisses the fears of some mathematicians that computers will lure them away from the methods that have served them so well for so long. "They're just afraid of change," he says mildly.

The Geometry Center is a hotbed of what is known as experimental mathematics, in which investigators test their ideas by representing them graphically and doing calculations on computers. Last year some of the center's faculty helped to found a journal, *Experimental Mathematics,* that showcases such work. "Experimental methods are not a new thing in mathematics," observes the journal's editor, David B. A. Epstein of the University of Warwick in England, noting that Carl Friedrich Gauss and other giants often performed experimental calculations before constructing formal proofs. "What's new is that it's respectable." Epstein acknowledges that not all his co-workers are so accepting. "One of my colleagues said, 'Your journal should be called the *Journal of Unproved Theorems.*' "

### Bubbles and Tortellini

A mathematician who epitomizes the new style of mathematics is Jean E. Taylor of Rutgers University. "The idea that you don't use computers is going to be increasingly foreign to the next generation," she says. For two decades, Taylor has investigated minimal surfaces, which represent the smallest possible area or volume bounded by a curve or surface. Perhaps the most elegant and simple minimal surfaces found in nature are soap bubbles and films. Taylor has always had an experimental bent. Early in her career she tested her handwritten models of minimal surfaces by dunking loops of wire into a sink of soapy water.

Now she is more likely to model bubbles with a sophisticated computer graphics program. She has also graduated from soap bubbles to crystals, which conform to somewhat more complicated rules about minimal surfaces. Together with Frederick J. Almgren of Princeton and Robert F. Almgren of the University of Chicago (her husband and stepson, respectively) and Andrew R. Roosen of the National Institute of Standards and Technology, Taylor is trying to mimic the growth of snowflakes and other crystals on a computer. Increasingly, she is collaborating with materials scientists and physicists, swapping mathematical ideas and programming techniques in exchange for clues about how real crystals grow.

Another mathematician who has prowled cyberspace in search of novel minimal surfaces is David A. Hoffman of the University of Massachusetts at Amherst. Among his favorite quarry are catenoids and helicoids, which resemble the pasta known as tortellini and were first discovered in the 18th century. "We gain a tremendous amount of intuition by looking at images of these surfaces on computers," he says.

In 1992 Hoffman, Fusheng Wei of Amherst and Hermann Karcher of the University of Bonn speculated on the existence of a new class of helicoids, ones with handles. They succeeded in representing these helicoids—the first discovered since the 18th century—on a computer and went on to produce a formal proof of their existence. "Had we not been able to see a picture that roughly corresponded to what we believed, we would never have been able to do it," Hoffman says.

The area of experimental mathematics that has received the lion's share of attention over the past decade is known as nonlinear dynamics or, more popularly, chaos. In general, nonlinear systems are governed by a set of simple rules that, through feedback and related effects, give rise to complicated phenomena. Nonlinear systems were investigated in the precomputer era, but computers allow mathematicians to explore these systems and watch them evolve in ways that Henri Poincaré and other pioneers of this branch of mathematics could not.

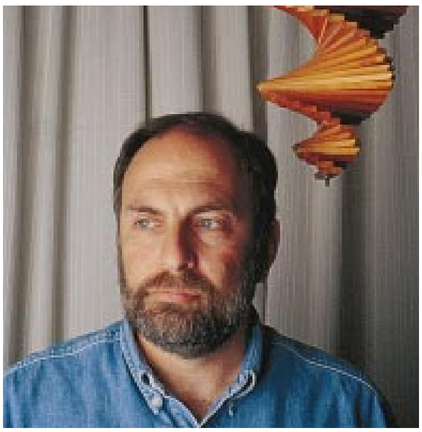
Cellular automata, which divide a computer screen into a set of cells (equivalent to pixels), provide a particularly dramatic illustration of the principles of nonlinearity. In general, the color, or "state," of each cell is determined by the state of its neighbors. A change in the state of a single cell triggers a cascade of changes throughout the system.

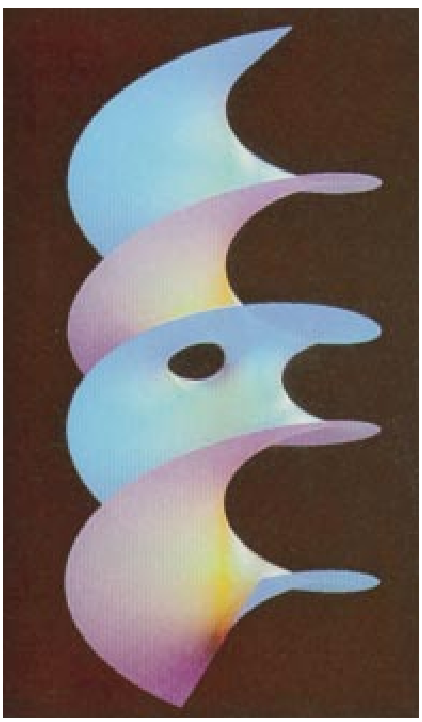
One of the most celebrated of cellular automata was invented by John H. Conway of Princeton in the early 1970s. Conway has proved that his automaton, which he calls "Life," is "undecidable": one cannot determine whether its patterns are endlessly variegated or eventually repeat themselves. Scientists have seized on cellular automata as tools for studying the origin and evolution of life. The computer scientist and physicist Edward Fredkin of Boston University has even argued that the entire universe is a cellular automaton.

More famous still is the Mandelbrot set, whose image has become an icon for the entire field of chaos since it was popularized in the early 1980s by Benoit B. Mandelbrot of the IBM Thomas J. Watson Research Center. The set stems from a simple equation containing a complex term (based on the square root of a negative number). The equation spits out solutions, which are then iterated, or fed back, into the equation.

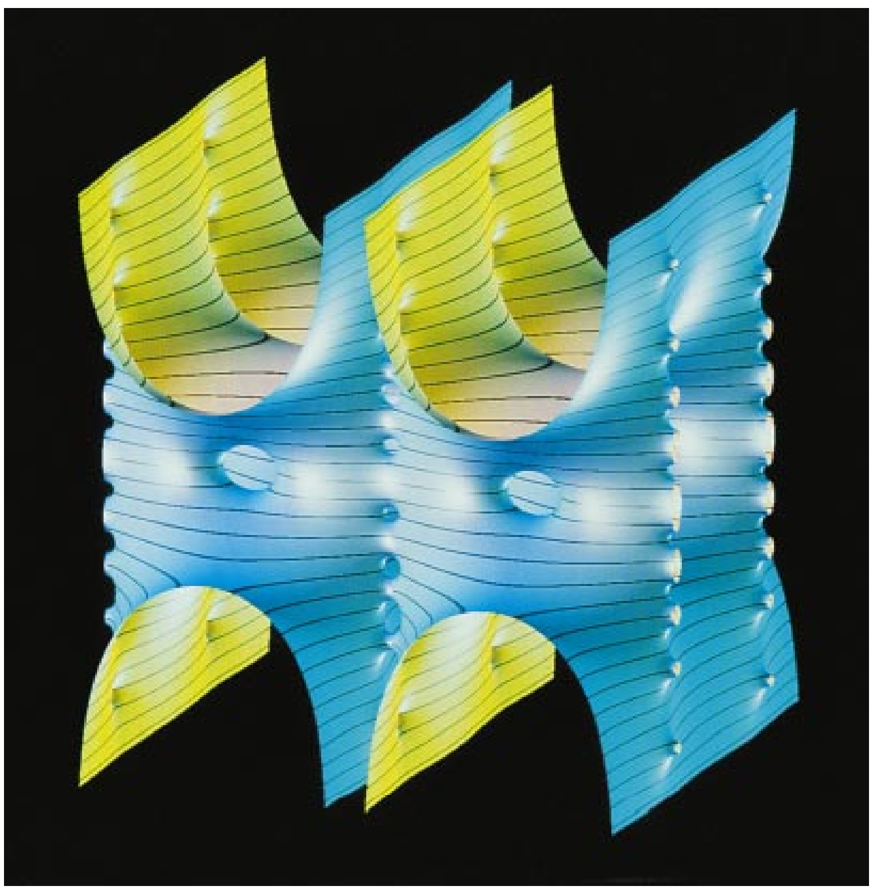
The mathematics underlying the set had been invented more than 70 years ago by two Frenchmen, Gaston Julia and Pierre Fatou, but computers laid bare their baroque beauty for all to see. When plotted on a computer, the Mandelbrot set coalesces into an image that has been likened to a tumorous heart, a badly burned chicken and a warty snowman. The image is a fractal: its fuzzy borders are infinitely long, and it displays patterns that recur at different scales.

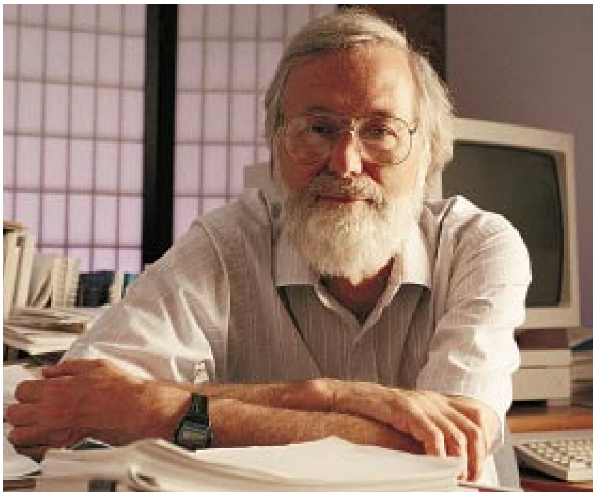
Researchers are now studying sets that are similar to the Mandelbrot set but inhabit four dimensions. "The kinds of complications you get here are the kinds you get in many different sciences," says John Milnor of the State University of New York at Stony Brook. Milnor is trying to fathom the properties of the four-dimensional set by examining two-dimensional slices of it generated by a computer. His preliminary findings led off the inaugural issue of *Experimental Mathematics* last year. Milnor, a 1962 Fields Medalist, says he

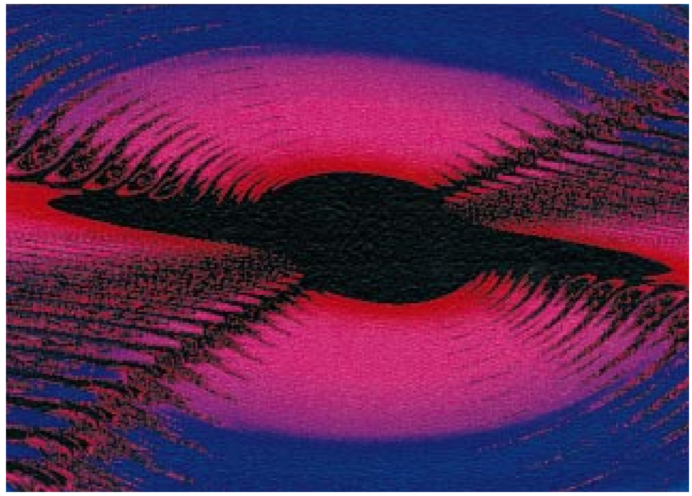




**HELICOID WITH A HOLE (*bottom left*) was discovered last year by David A. Hoffman of the University of Massachusetts at Amherst and his colleagues, with the help of computer graphics. Edward C. Thayer, one of Hoffman's graduate students, recently found a structure (*below*) that mimics the pattern of certain polymers.**





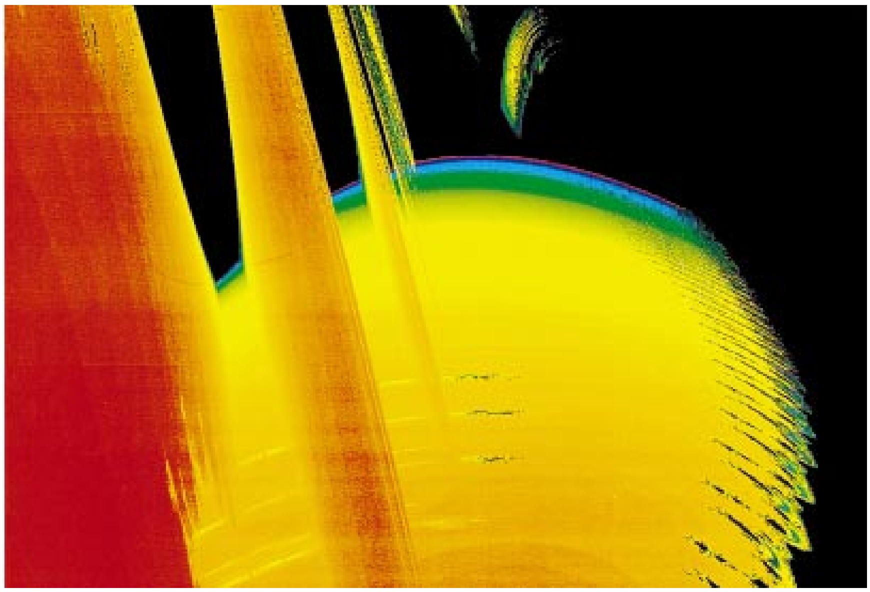


occasionally performed computer experiments in the days of punch cards, but "it was a miserable process. It has become much easier."

The popularity of graphics-oriented mathematics has provoked a backlash. Krantz of Washington University charged four years ago in the *Mathematical Intelligencer* that "in some circles, it is easier to obtain funding to buy hardware to generate pictures of fractals than to obtain funding to study algebraic geometry."

A broader warning about "speculative" mathematics was voiced this past July in the *Bulletin of the American Mathematical Society* by Arthur Jaffe of Harvard and Frank S. Quinn of the Virginia Polytechnic Institute. They suggested that computer experiments and correspondence with natural phenomena are no substitute for proofs in establishing truth. "Groups and individuals within the mathematics community have from time to time tried being less compulsive about details of arguments," Jaffe and Quinn wrote. "The results have been mixed, and they have occasionally been disastrous."

Most mathematicians exploiting computer graphics and other experimental techniques agree that seeing should not be believing and that proofs are still needed to verify the conjectures they arrive at through computation. "I think mathematicians were contemplating their navels for too long, but that doesn't mean I think proofs are irrelevant," Taylor says. Hoffman offers an even stronger defense of traditional proofs. "Proofs are the only laboratory instrument mathematicians have," he remarks, "and they are in danger of being thrown out." Although computer graphics are "unbelievably wonderful," he adds, "in the 1960s drugs were unbelievably wonderful, and some people didn't survive."



**UNEARTHLY LANDSCAPES emerge when a computer generates "slices" of a four-dimensional map similar to the well-known Mandelbrot set. John Milnor of the State University of New York at Stony Brook studies similar two-dimensional images in order to understand the properties of the complex mathematical object.**

Indeed, veteran computer enthusiasts know better than most that computational experiments—whether involving graphics or numerical calculations—can be deceiving. One cautionary tale involves the Riemann hypothesis, a famous prediction about the patterns displayed by prime numbers as they march toward infinity. First posed more than 100 years ago by Bernhard Riemann, the hypothesis is considered to be one of the most important unsolved problems in mathematics.

A contemporary of Riemann's, Franz Mertens, proposed a related conjecture involving positive whole numbers; if true, the conjecture would have provided strong evidence that the Riemann hypothesis was also true. By the early 1980s computers had shown that Mertens's proposal did indeed hold for at least the first 10 billion integers. In 1984, however, more extensive computations revealed that eventually—at numbers as high as $10^{(10^{70})}$—the pattern predicted by Mertens vanishes.

One potential drawback of computers is that all their calculations are based on the manipulation of discrete, whole numbers—in fact, ones and zeros. Computers can only approximate real numbers, such as *pi* or the square root of two. Someone knowledgeable about the rounding-off functions of a simple pocket calculator can easily induce it to generate incorrect answers to calculations. More sophisticated programs can make more complicated and elusive errors. In 1991 David R. Stoutemyer, a software specialist at the University of Hawaii, presented 18 experiments in algebra that gave wrong answers when performed with standard mathematics software.

Stephen Smale of the University of California at Berkeley, a 1966 Fields Medalist, has sought to place mathematical computation on a more secure foundation—or at least to point out the size and location of the cracks running through the foundation. Together with Lenore Blum of the Mathematical Sciences Research Institute at Berkeley and Michael Shub of IBM, he has created a theoretical model of a computer that can process real numbers rather than just integers.

Blum and Smale recently concluded that the Mandelbrot set is, in a technical sense, uncomputable. That is, one cannot determine with certainty whether any given point on the complex plane resides within or outside the set's hirsute border. These results suggest that "you have to be careful" in extrapolating from the results of computer experiments, Smale says.

These concerns are dismissed by Stephen Wolfram, a mathematical physicist at the University of Illinois. Wolfram is the creator of Mathematica, which has become the leading mathematics software since first being marketed five years ago. He acknowledges that "there are indeed pitfalls in experimental mathematics. As in all other kinds of experiments, you can do them wrong." But he emphasizes that computational experiments, intelligently performed and analyzed, can yield more results than the old-fashioned conjecture-proof method. "In every other field of science there are a lot more experimentalists than theorists," Wolfram says. "I suspect that will increasingly be the case with mathematics."

"The obsession with proof," Wolfram declares, has kept mathematicians from discovering the vast new realms of phenomena accessible to computers. Even the most intrepid mathematical experimentalists are for the most part "not going far enough," he says. "They're taking existing questions in mathematics and investigating those. They are adding a few little curlicues to the top of a gigantic structure."

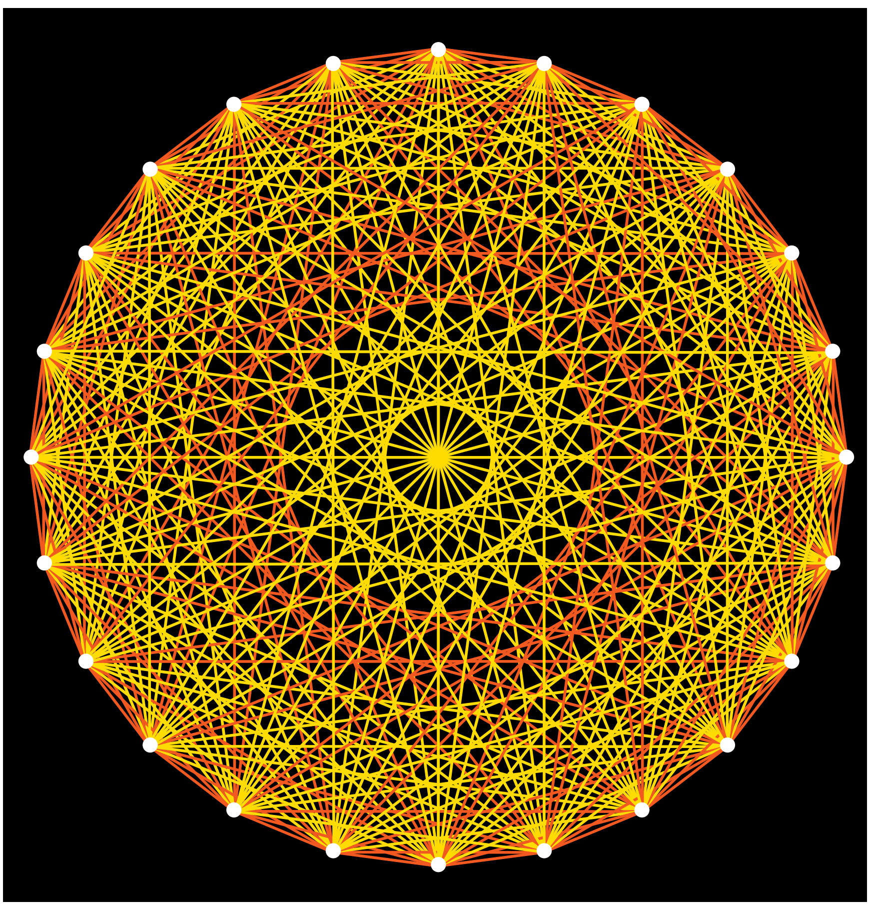
Mathematicians may take this view with a grain of salt. Although he shares Wolfram's fascination with cellular automata, Conway contends that Wolfram's career—as well as his contempt for proofs—shows he is not a real mathematician. "Pure mathematicians usually don't found companies and deal with the world in an aggressive way," Life's creator says. "We sit in our ivory towers and think about things."

Purists may have a harder time ignoring William P. Thurston, who is also an enthusiastic booster of experimental mathematics and of computers in mathematics. Thurston, who heads the Mathematical Sciences Research Institute at Berkeley and is a co-director of the Geometry Center (with Albert Marden of the University of Minnesota), has impeccable credentials. In the mid-1970s he pointed out a deep potential connection between two separate branches of mathematics—topology and geometry. Thurston won a Fields Medal for this work in 1982.

Thurston emphasizes that he believes mathematical truths are discovered and not invented. But on the subject of proofs, he sounds less like a disciple of Plato than of Thomas S. Kuhn, the philosopher who argued in his 1962 book, *The Structure of Scientific Revolutions,* that scientific theories are accepted for social reasons rather than because they are in any objective sense "true." "That mathematics reduces in principle to formal proofs is a shaky idea" peculiar to this century, Thurston asserts. "In practice, mathematicians prove theorems in a social context," he says. "It is a socially conditioned body of knowledge and techniques."

The logician Kurt Gödel demonstrated more than 60 years ago through his incompleteness theorem that "it is impossible to codify mathematics," Thurston notes. Any set of axioms yields statements that are self-evidently true but cannot be demonstrated with those axioms. Bertrand Russell pointed out even earlier that set theory, which is the basis of much of mathematics, is rife with logical contradictions related to the problem of self-reference. (The self-contradicting statement "This sentence is false" illustrates the problem.) "Set theory is based on polite lies, things we agree on even though we know they're not true," Thurston says. "In some ways, the foundation of mathematics has an air of unreality."

Thurston thinks highly formal proofs are more likely to be flawed than those appealing to a more intuitive level of understanding. He is particularly enamored of the ability of computer graphics



**PARTY PROBLEM was solved after a vast computation by Stanislaw P. Radziszowski and Brendan D. McKay. They calculated that at least 25 people are required to ensure either that four people are all mutual acquaintances or that five are mutual strangers. This diagram, in which red lines connect friends and yellow lines link strangers, shows that a party of 24 violates the dictum.**

## Silicon Mathematicians

The continuing penetration of computers into mathematics has revived an old debate: Can mathematics be entirely automated? Will the great mathematicians of the next century be made of silicon?

In fact, computer scientists have been working for decades on programs that generate mathematical conjectures and proofs. In the late 1950s the artificial-intelligence guru Marvin Minsky showed how a computer could "rediscover" some of Euclid's basic theorems in geometry. In the 1970's Douglas Lenat, a former student of Minsky's, presented a program that devised even more advanced geometry theorems. Skeptics contended that the results were, in effect, embedded in the original program.

A decade ago the computer scientist and entrepreneur Edward Fredkin sought to revive the sagging interest in machine mathematics by creating what came to be known as the Leibniz Prize. The prize, administered by Carnegie Mellon University, offers $100,000 for the first computer program to devise a theorem that has a "profound effect" on mathematics.

Some practitioners of what is known as automated reasoning think they may be ready to claim the prize. One is Larry Wos of Argonne National Laboratory, editor of the *Journal of Automated Reasoning.* He claims to have developed a program that has solved problems in mathematics and logic "that have stumped people for years." Another is Siemeon Fajtlowicz of the University of Houston, inventor of a program, called Graffiti, that has proposed "thousands" of conjectures in graph theory.

None of these achievements comes close to satisfying the "profound effect" criterion, according to David Mumford of Harvard University, a judge for the prize. "Not now, not 100 years from now," Mumford replies when asked to predict when the prize might be claimed.

Some observers think computers will eventually surpass our mathematical abilities. After all, notes Ronald L. Graham of AT&T Bell Laboratories, "we're not very well adapted for thinking about the space-time continuum or the Riemann hypothesis. We're designed for picking berries or avoiding being eaten."

Others side with the mathematical physicist Roger Penrose of the University of Oxford, who in his 1989 book, *The Emperor's New Mind,* asserted that computers can never replace mathematicians. Penrose's argument drew on quantum theory and Gödel's incompleteness theorem, but he may have been most convincing when discussing his personal experience. At its best, he suggested, mathematics is an art, a creative act, that cannot be reduced to logic any more than *King Lear* or Beethoven's Fifth can.

to communicate abstract mathematical concepts to others both within and outside the professional community. Two years ago, at his urging, the Geometry Center produced a computer-generated "video proof," called Not Knot, that dramatizes a ground-breaking conjecture he proved a decade ago [*see illustration on pages 92 and 93*]. Thurston mentions proudly that the rock band the Grateful Dead has shown the Not Knot video at its concerts.

Whether Deadheads grok the substance of the video—which concerns how mathematical objects called three-manifolds behave in a non-Euclidean "hyperbolic" space—is another matter. Thurston concedes that the video is difficult for nonmathematicians, and even some professionals, to fathom, but he is undaunted. The Geometry Center is now producing a video of yet another of his theorems, which demonstrates how a sphere can be turned inside out [*see cover illustration*]. Last fall, moreover, Thurston organized a workshop at which participants discussed how virtual reality and other advanced technologies could be adapted for mathematical visualization.

Paradoxically, computers have catalyzed a countertrend in which truth is obtained at the expense of comprehensibility. In 1976 Kenneth Appel and Wolfgang Haken of the University of Illinois claimed they had proved the four-color conjecture, which stated that four hues are sufficient to construct even an infinitely broad map so that no identically colored countries share a border. In some respects, the proof of Appel and Haken was conventional—that is, it consisted of a series of logical, traceable steps proceeding to a conclusion. The conclusion was that the conjecture could be reduced to a prediction about the behavior of some 2,000 different maps.

Since checking this prediction by hand would be prohibitively time-consuming, Appel and Haken programmed a computer to do the job for them. Some 1,000 hours of computing time later, the machine concluded that the 2,000 maps behave as expected: the four-color conjecture was true.

### The Party Problem

Other computer-assisted proofs have followed. Just this year, a proof of the so-called party problem was announced by Stanislaw P. Radziszowski of the Rochester Institute of Technology and Brendan D. McKay of the Australian National University in Canberra. The problem, which derives from work in set theory by the British mathematician Frank P. Ramsey in the 1920s, can be phrased as a question about relationships between people at a party. What is the minimum number of guests that must be invited to guarantee that at least $X$ people are all mutual acquaintances or at least $Y$ are mutual strangers? This number is known as a Ramsey number.

Previous proofs had established that 18 guests are required to ensure that there are either four mutual acquaintances or four strangers. In their proof, Radziszowski and McKay showed that the Ramsey number for four friends or five strangers is 25. Socialites might think twice about trying to calculate the Ramsey number for greater $X$'s and $Y$'s. Radziszowski and McKay estimate that their proof consumed the equivalent of 11 years of computation by a standard desktop machine. That may be a record, Radziszowski says, for a problem in pure mathematics.

The value of this work has been debated in an unlikely forum—the newspaper column of advice-dispenser Ann Landers. In June a correspondent complained to Landers that resources spent on the party problem should have been used to help "starving children in war-torn countries around the world." Some mathematicians raise another objection to computer-assisted proofs. "I don't believe in a proof done by a computer, says Pierre Deligne of the Institute for Advanced Study, an algebraic geometer and 1978 Fields Medalist. "In a way, I am very egocentric. I believe in a proof if I understand it, if it's clear." While recognizing that humans can make mistakes, he adds: "A computer will also make mistakes, but they are much more difficult to find."

Others take a more functional point of view, arguing that establishing truth is more important than giving mathematicians an aesthetic glow, particularly if a result is ever to find an application. Defenders of this approach, who tend to be computer scientists, point out that conventional proofs are far from immune to error. At the turn of the century, most theorems were short enough to read in one sitting and were produced by a single author. Now proofs often extend to hundreds of pages or more and are so complicated that years may pass before they are confirmed by others.

The current record holder of all conventional proofs was completed in the early 1980s and is called the classification of finite, simple groups. (A group is a set of elements, such as integers, together with an operation, such as addition, that combines two elements to get a third one.) The demonstration consists of some 500 articles totaling nearly 15,000 pages and written by more than 100 workers. It has been said that the only person who grasped the entire proof was its general contractor, Daniel Gorenstein of Rutgers. Gorenstein died last year.

Much shorter proofs can also raise doubts. Three years ago Wu-Yi Hsiang of Berkeley announced he had proved an old conjecture that one can pack the most spheres in a given volume by stacking them like cannonballs. Today some skeptics are convinced the 100-page proof is flawed; others are equally certain it is basically correct.

Indeed, the key to greater reliability, according to some computer scientists, is not less computerization but more. Robert S. Boyer of the University of Texas at Austin has led an effort to squeeze the entire sprawling corpus of modern mathematics into a single data base whose consistency can be verified through automated "proof checkers."

The manifesto of the so-called QED Project states that such a data base will enable users to "scan the entirety of mathematical knowledge for relevant results and, using tools of the QED system, build upon such results with reliability and confidence but without the need for minute comprehension of the details or even the ultimate foundations." The QED system, the manifesto proclaims rather grandly, can even "provide some antidote to the degenerative effects of cultural relativism and nihilism" and, presumably, protect mathematics from the all-too-human willingness to succumb to fashion.

The debate over computer proofs has intensified recently with the advent of a technique that offers not certainty but only a statistical probability of truth. Such proofs exploit methods similar to those underlying error-correction codes, which ensure that transmitted messages are not lost to noise and other effects by making them highly redundant. The proof must first be spelled out precisely in a rigorous form of mathematical logic. The logic then undergoes a further transformation called arithmetization, in which "and," "or" and other functions are translated into arithmetic operations, such as addition and multiplication.

Like a message transformed by an error-correction code, the "answer" of a probabilistic demonstration is distributed throughout its length—as are any errors. One checks the proof by querying it at different points and determining whether the answers are consistent; as the number of checks increases, so does the certainty that the argument is correct. Laszlo Babai of the University of Chicago, who developed the proofs two years ago (along with Lance Fortnow, Carsten Lund and Mario Szegedy of Chicago and Leonid A. Levin of Boston University), calls them "transparent." Manuel Blum of Berkeley, whose work helped to pave the way for Babai's group, suggests the term "holographic."

### The Uncertain Future

Whatever they are named, such proofs have practical drawbacks. Szegedy acknowledges that transforming a conventional demonstration into the probabilistic form is difficult, and the result can be a "much bigger and uglier animal." A 1,000-line proof, for example, could easily balloon to $1{,}000^3$ (1,000,000,000) lines. Yet Szegedy contends that if he and his colleagues can simplify the transformation process, probabilistic proofs might become a useful method for verifying mathematical propositions and large computations—such as those leading to the four-color theorem. "The philosophical cost of this efficient method is that we lose the absolute certainty of a Euclidean proof," Babai noted in a recent essay. "But if you do have doubts, will you bet with me?"

Such a bet would be ill advised, Levin believes, since a relatively few checks can make the chance of error vanishingly small: one divided by the number of particles in the universe. Even the most straightforward conventional proofs, Levin points out, are susceptible to doubts of this scale. "At the moment you find an error, your brain may disappear because of the Heisenberg uncertainty principle and be replaced by a new brain that thinks the proof is correct," he says.

Ronald L. Graham of AT&T Bell Laboratories suggests that the trend away from short, clear, conventional proofs that are beyond reasonable doubt may be inevitable. "The things you can prove may be just tiny islands, exceptions, compared to the vast sea of results that cannot be proved by human thought alone," he explains. Mathematicians seeking to navigate uncharted waters may become increasingly dependent on experiments, probabilistic proofs and other guides. "You may not be able to provide proofs in a classical sense," Graham says.

Of course, mathematics may yield fewer aesthetic satisfactions as investigators become more dependent on computers. "It would be very discouraging," Graham remarks, "if somewhere down the line you could ask a computer if the Riemann hypothesis is correct and it said, 'Yes, it is true, but you won't be able to understand the proof.' "

Traditionalists no doubt shudder at the thought. For now, at least, they can rally behind heros like Wiles, the conqueror of Fermat's last theorem, who eschews computers, applications and other abominations. But there may be fewer Wileses in the future if reports from the front of precollege education are any guide. The Mathematical Sciences Research Institute at Berkeley, which is overseen by Thurston, has been holding an ongoing series of seminars with high school teachers to find new ways to entice students into mathematics. This past January Lenore Blum, the institute's deputy director, organized a seminar devoted to the question "Are Proofs in High School Geometry Obsolete?"

The mathematicians insisted that proofs are crucial to ensure that a result is true. The high school teachers demurred, pointing out that students no longer considered traditional, axiomatic proofs to be as convincing as, say, visual arguments. "The high school teachers overwhelmingly declared that most students now (Nintendo/joystick/MTV generation) do not relate to or see the importance of 'proofs,' " the minutes of the meeting stated. Note the quotation marks around the word "proofs."

FURTHER READING


ISLANDS OF TRUTH: A MATHEMATICAL MYSTERY CRUISE. Ivars Peterson. W. H. Freeman and Company, 1990.

THE PROBLEMS OF MATHEMATICS. Ian Stewart. Oxford University Press, 1992.

PI IN THE SKY: COUNTING, THINKING, AND BEING. John D. Barrow. Oxford University Press, 1992.

## CRADA Mania

*Will joint research make U.S. industry competitive?*

When economic competitiveness became a front-burner issue during the 1980s, Congress passed laws to entice companies to adopt technologies developed in government laboratories and encourage government researchers to team up with industry. The Clinton administration, keen to secure benefits from research, is going even further down that road. It is telling the Department of Energy's huge national laboratories, as well as the National Institutes of Health and other agencies, that they must make their expertise available to industry. But change is bringing problems in its wake.

To show that they are solidly on board, officials and laboratory directors point proudly to the explosion in cooperative research and development agreements, or CRADAs. These agreements between government laboratories and industry or universities allow participants to share the costs of a collaboration. In return, the nongovernment partner is granted some rights over any offspring, in the form of intellectual property. The national laboratories have been allowed to have CRADAs only since 1989, but they now boast close to 500. Their directors talk expansively of devoting 20 percent of their budgets to CRADAs and like arrangements within a few years.

RANDY MONTOYA *Sandia National Laboratories*

***ENERGY SECRETARY Hazel R. O'Leary tours Sandia National Laboratories in April after signing a CRADA with Sematech. With her is Senator Jeff Bingaman of New Mexico.***

Opinions on the value of CRADAs vary widely in industry and even within the government, despite their mushrooming growth. Some critics in Congress say CRADAs can easily turn into a fire sale of taxpayer-supported research. And companies complain that even with the new laws, red tape means it is still too hard to collaborate with researchers in government laboratories. "In many ways the jury is still out on these arrangements," says Thomas H. Moss, director of graduate studies and research at Case Western Reserve University in Cleveland. "It's easy to imagine that people in government are moving to achieve the maximum number of CRADAs independently of substance."

Indeed, the rush into CRADAs has been haphazard. The 1986 law authorizing them was not accompanied by guidelines. The result was a profusion of CRADAs with widely different terms. The NIH, for one, has been changing its policies. Dinah Singer, an immunologist who chaired a committee that drafted CRADA guidelines for the NIH, complains that some of the health institutes' early CRADAs failed to define clearly enough the area they covered. "It's clear research scope should be limited to where an invention has been made in a federal laboratory and development is to be transferred to the private sector. It's not appropriate for a CRADA to apply to concepts or ideas," she says.

CRADAs with drug companies are particularly controversial. A major unresolved question for the NIH is how—or whether—to require companies that develop drugs based on an NIH discovery to sell them at a "fair" price. Pharmaceutical companies refuse to divulge how they set their prices. Their reticence prompted Congressman Ronald L. Wyden of Oregon, a longtime scourge of the industry, to demand to know why the NIH had not limited the price that Bristol-Myers Squibb charges for Taxol, a drug whose anticancer properties were discovered by the National Cancer Institute.

The inspector general of the Department of Health and Human Services, in a report now under review at the department, also finds cause for concern that the NIH's CRADAs do not protect the public interest. Fears that researchers could be left with divided loyalties have prompted an interagency committee under the Department of Commerce to draw up conflict-of-interest guidelines for federal employees in CRADAs.

The most common complaint from academic researchers and industry is the delay engendered by legal review of agreements. The story of Stephen M. Boyle, a researcher at the Virginia-Mary-land Regional College of Veterinary Medicine who has a CRADA with the Walter Reed Army Institute of Research, is typical. Boyle says that although it took him and his collaborators at Walter Reed only an hour to agree on a one-page statement covering their work on a human brucellosis vaccine, it took lawyers six months to approve it. The formal document, Boyle recounts ruefully, is eight pages long and has the work statement attached at the end.

The General Accounting Office reported recently that the Department of Energy takes between three and five times longer to sign off on a CRADA, on average, than either the army or the National Institute of Standards and Technology. At the DOE, eight months has been the average, and a year is not unusual. One reason, according to James E. Wells, Jr., an associate director for energy and science issues at the accounting office, is "the desire of some officials to create a highly visible, separately funded CRADA program in order to justify continued support for the department's weapons laboratories." In addition, the DOE has a separate fund for CRADAs, so it has to have a formal competitive selection and approval process. The National Institute of Standards and Technology, in contrast, which has less formal procedures, has been able to approve CRADAs in as little as two weeks.

Another reason for the Department of Energy's slowness is that its national laboratories are managed by independent contractors. That means agreements have to be monitored by DOE headquarters in Washington, D.C. Secretary of Energy Hazel R. O'Leary told Congress in late July that the department will shortly produce a simplified, approved model CRADA for small businesses. Moreover, by November laboratory directors will be empowered to authorize CRADAs that involve expenditures of less than $500,000. O'Leary said she will also cut the time for approving a CRADA at DOE headquarters from 32 weeks to fewer than 16.

Delays are not industry's only complaint. Because the DOE attracts about 10 times more CRADA proposals than it can approve, many companies expend substantial efforts drafting and negotiating a proposal only to have it turned down months later, notes John L. Sullivan, software counsel of Cray Research, Inc. Roger G. Little, chairman of the Solar Energy Industries Association, told a House subcommittee in June that the requirement that industrial partners match the value of government contributions "keeps CRADAs from providing real working capital for small businesses."

Large companies have their own gripes. The Department of Energy has clashed with General Motors and the Ford Motor Company over the DOE's policies on product liability and a "U.S. preference" provision. Until recently the DOE required that products developed from CRADA research be substantially manufactured in the U.S. or be otherwise beneficial to the domestic economy. But global companies—the principal users of DOE-developed technology—find these and other restrictions overly burdensome, says John J. Sheridan, director of technology sourcing at the National Center for Manufacturing Sciences in Ann Arbor, Mich. The DOE has recently indicated a willingness to relax the U.S. preference provision.

The DOE has also refused to offer indemnification against product liability arising from joint work beyond a narrow limit. But the limits the agency is proposing are unacceptable to industry, Sheridan asserts. Moreover, he says, industry views as "unreasonable" attempts by the DOE to specify in advance the permissible field of application of a technology developed in a CRADA. "Many of our industry members have concluded that CRADAs do not work and have turned their attention elsewhere," Sheridan told Congress.

He supports instead alternative routes

## A Wand for the Meter Reader

When gaslights were introduced into homes and businesses in the late 19th century, suppliers billed according to the number of lamps at the establishment. After heating and cooking made gas consumption more variable, the bellows meter was developed. A complicated mechanism, this instrument originally used a pair of inflatable sheep bladders linked by crankshaft and connecting rods to ratchet- and gear-driven dials.

Some 100 years later the bellows have been changed to rubber, but the same ungainly design is still in use all over the world. The mundane meter is the most common home appliance that is not directly purchased by consumers. More than 10 million are sold annually to gas utilities, supporting a $700-million-a-year market. But not for much longer. These unsightly cast-iron lawn ornaments may be supplanted by a new electronic version developed after five years of work by government boffins Down Under.

Researchers at Australia's Commonwealth Scientific and Industrial Research Organization (CSIRO), in a joint project with the utility AGL Australia, have come up with a smaller, less expensive and highly accurate alternative. The developers believe the new meters will last twice as long and reduce the potential margin of error from as high as 10 to 0.2 percent. The devices may help make the monthly visit of the meter reader a piece of nostalgia, like the hurdy-gurdy man coming down the block.

The CSIRO unit, a solid-state device about the size of a videotape cartridge, measures the velocity of an ultrasonic sound burst to determine gas flow. The 500-gram, battery-powered meter ricochets bursts of acoustical energy across a pair of transducers along a pipe 14 millimeters in diameter. The sound pulse travels faster with the gas flow than against it. The fluid velocity can be solved knowing just the upstream and downstream times along with the fixed distance between the detectors. A small, low-power custom microprocessor handles all the control, calculation and display functions.

The meter is projected to have a useful operating life of more than 25 years and requires a battery change once every eight years. The key to this long duty cycle was the development of a highly efficient piezoelectric transducer that converts the vibration of the sound into an electric signal. The CSIRO meter can be read with an infrared wand, similar to a television remote control, at a distance of three meters.

CSIRO is not alone. For the past 10 years, gas companies throughout the world have sponsored efforts to develop a more compact electronic system. In 1987 British Gas held a competition for a new design, for which it received 32 submissions. It chose to fund development projects from Siemens, the German electronics giant, and Gill, a small British start-up company. Their systems are still shrouded in confidentiality. Meanwhile meter readers in California will be waving wands this autumn in a test of the CSIRO meter.

*—Joshua Shapiro*

for technology transfer, such as direct funding of industrial research consortia. Yet while the NIH is seeking to restrict CRADAs to narrow fields of work, the DOE is moving in the opposite direction, seeking larger CRADAs that involve industrial consortia.

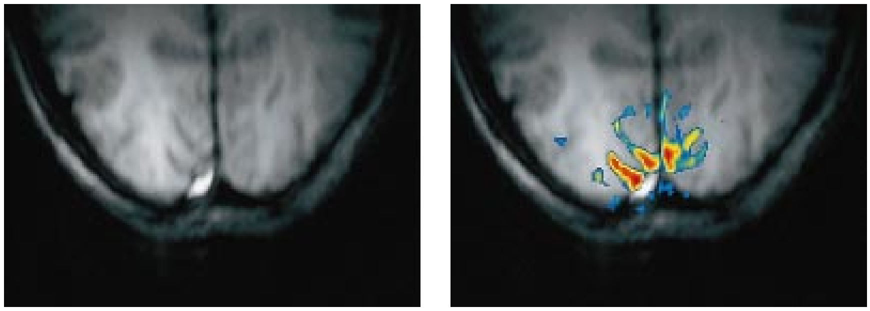
The conclusion that CRADAs have been less than a resounding success is supported by a study conducted by J. David Roessner and Anne F. Wise of the Georgia Institute of Technology. The study, which focused on industrial and federal laboratories, concluded that the proportion of research-intensive companies that have contacts with federal laboratories was unchanged over the past five years. In other words, some companies are interacting more now than they were—but the pool of industrial collaborators is not expanding.

Surprisingly, the companies in Roessner and Wise's report that did work with federal laboratories said that expectations of licensing technology were not the main reason for their forging links. Rather, it was access to the unique resources of the laboratories. According to Roessner, technology-based companies still see federal laboratories as relatively unimportant sources of new ideas, in comparison with universities and other companies. This was not the result that the legislation encouraging CRADAs sought.

Congress, buffeted by the conflicting opinions, is now considering additional measures to improve technology transfer in general and CRADAs in particular. Legislation is pending in both the House and the Senate; most of the efforts focus on making it easier for industry to find a federal partner. But easier access could give rise to more problems unless care is taken to avoid the sort of situation that the not-for-profit Scripps Research Institute in La Jolla, Calif., found itself in earlier this year.

Scripps was about to sign a large research agreement with Sandoz, the Swiss-based pharmaceutical company, when word leaked out. For $300 million over 10 years, the agreement would have given the drug company rights to the commercial fruits of federally supported research at Scripps. The deal was denounced by the then director of the NIH, Bernadine Healy, as giving Sandoz "excessive control," and it is now being renegotiated.

The day when Los Alamos and Sandia National Laboratories might be working principally for an industrial conglomerate is probably still far off. But their directors had better be ready to face tough questions as industrial influence at the laboratories grows. *—Tim Beardsley*



KAMIL UGURBIL University of Minnesota

***HIGH-STRENGTH MRI SCANS from the University of Minnesota Medical School show areas of cortex both before* (left) *and during* (right) *stimulation by flashing lights.***

## Magnetic Apprehensions

*Radiologists call for more testing of MRI's effects*

A new generation of fast and very powerful magnetic resonance imaging (MRI) machines is providing scientists with an unparalleled view not just of tissues but also of the metabolic changes associated with functioning brains and other organs. Recent investigations have shown the activity in specific areas of the brain during thought, conversation or dreaming. But even as the exciting findings are reported, a few radiologists are voicing concern that another round of testing is needed to evaluate biological effects and safety issues for these advanced scanners. "A lot of the work should be repeated," says Frank Shellock, a radiologist and co-chairman of the safety committee of the Society of Magnetic Resonance Imaging.

Shellock is concerned because to take rapid snapshots of physiological processes, the rate that scanner fields are pulsed through the body has been accelerated in the new machines, and the magnetic field, already 30,000 times that of the earth, has in some cases been increased by a factor of three or four. "I don't feel you can look at the data accumulated at the lower field strengths and at different radio frequencies and assume that the same safety issues have been addressed at high fields," Shellock notes.

Most scientists and radiologists are comfortable with the safety of the lower-speed machines, with fields of about one to 1.5 teslas, now in clinical use—in fact, they are widely viewed as a benign alternative to x-rays. "There have been thousands of people scanned, and the identifiable problems are mostly failure of the operator or the machine in one form or another," says Robert Balaban, chief of the laboratory of cardiac energetics at the National Institutes of Health.

But many of the safety studies used machines of even lower field strengths or were conducted on cell cultures or animals, not people. Although functional imaging has been demonstrated on 1.5-tesla machines retrofitted with hardware to allow faster scanning, scientists are looking to higher strength machines to provide the image resolution they seek. There are only half a dozen or so experimental machines in the U.S. operating in the neighborhood of four teslas, but more are being built.

Lawrence Berkeley Laboratory wants to build a 10-tesla machine. That field strength exceeds that of the six-tesla magnets in the proposed Superconducting Super Collider.

Assessing MRI safety has proved difficult because the devices produce an exotic combination of three disparate electromagnetic fields. In addition to the unchanging or static magnetic field, another magnetic field, called a gradient field, increases in intensity across the body to provide the scanner with information about the location of the tissues being imaged. Furthermore, coils emit rapid pulses of electromagnetic energy at radio frequencies to induce the hydrogen nuclei in the body to move out of alignment with the static field. The measurement of this misalignment yields an image.

The initial focus of safety testing was radio-frequency energy, because scientists have long known that it can heat body tissues, although the small temperature rise caused by present MRI machines has been judged of no real concern. Shellock, however, worries that different radio-frequency pulse sequences designed to improve the image contrast and to shorten the time between images for "functional" imaging could induce significant heating. For a healthy person, the temperature rise would be similar to sitting at length in the sun on a hot day. But Shellock says ill or sedated patients with poor temperature regulation "may not be able to withstand those temperatures."

The new machines speed imaging primarily by switching gradient fields faster. But the rapid fluctuation of field strengths has the potential to stimulate nerves. At the threshold of stimulation, people begin to experience sensations such as a light tapping on the nose, the small of the back or the inside of the thigh. A subject in functional imaging experiments at Lawrence Berkeley Laboratory reported that at levels not much higher than the threshold, the feeling resembled an electric shock to the abdomen.

Despite these prospects, many radiologists believe that rapidly switched gradients are not harmful, a position with which the Food and Drug Administration concurs. The FDA has recently allowed an exception to its guidelines, so some machines now in clinical use operate just below the threshold of stimulation. “Peripheral nerve stimulation can serve as an early warning to prevent cardiac stimulation and is not in itself harmful,” says Whit Athey, a senior FDA scientist. But the static magnetic field in the high-field, four-tesla machines has caused a small percentage of volunteers to report feeling nauseated, seeing flashing lights or tasting metal in their mouths—sensations that scanner manufacturers say can be reduced by having a patient avoid moving his or her head during an examination that may last about half an hour.

To be certain these effects are benign, Shellock and others are calling for more scrutiny. The only conference to date devoted solely to the safety of MRI was held in 1991 by the New York Academy of Sciences. It drew a rare congregation of radiologists, biologists and electrical engineers. “I was struck by how little knowledge or widespread agreement there is about the biological effects,” says radiologist Jeffrey Weinreb, director of MRI at the New York University Medical Center, who was a session chairman. “The models and the numbers are not in.” *—Karen Fitzgerald*

## Mag Lift

*Japan’s engineers push the envelope for elevators*

Long before Chicago’s Sears Tower or New York’s World Trade Center was built, Frank Lloyd Wright conceived of a city contained in a single mile-high skyscraper. Wright’s dream never went further from his drawing board than the pages of coffee-table books on architecture. Had the structure been built, it would have been the world’s tallest walk-up. The hike would have been necessary because conventional elevator technology reaches its limits at about 800 meters. Above that height, elevator cables become too thick and too heavy for practical use.

Decades after Wright’s death in 1959, engineers in Japan with a similar penchant for futuristic indulgences are planning structures with similar spans—but

## Chronologically Privileged

C-SPAN viewers who watched Congress this spring tiptoe around the subject of Social Security cuts may have smiled to hear politically correct legislators refer deferentially to their elderly constituents as “the chronologically privileged.” One point of their debate is whether the senior citizenry is financially privileged as well. Yeasayers point to U.S. Census Bureau statistics that show a poverty rate since 1985 running about 10 percent lower for elders than for Americans as a whole. Naysayers question the bureau’s method, which sets a lower poverty level for retirees in order to compensate for reduced work-related expenses.

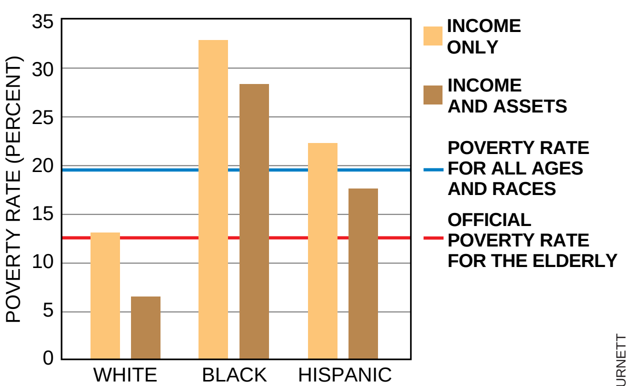
To help clarify the situation, the American Association of Retired Persons’ Andrus Foundation joined with the University of Wisconsin to commission Michael S. Rendall of Cornell University and Alden Speare, Jr., of Brown University to study poverty among the elderly. In addition to measuring income by traditional methods, the sociologists counted elders’ assets and government subsidies. The results may have come as a surprise. Despite raising the poverty level 25 percent above that used for official statistics—“it’s widely believed that the thresholds are kept down so that poverty doesn’t look as bad as it is,” Rendall explains—the researchers found that the poverty rate for those age 65 and older dropped from 15.1 to 8.9 percent.

It dropped further for some groups of seniors than others. “Blacks and other minorities tend to have lower assets than whites, and the gap in assets is much bigger than the gap in incomes,” Speare observes. Elderly white households owned median total assets worth $84,760, compared with a median of $12,121 for black elderly. Rendall and Speare found a poverty rate of 28.3 percent for black pensioners, 4.2 times that of elderly whites.

The sociologists are quick to point out that their figures depend on a critical—and questionable—assumption: that retirees will sell off their assets evenly and die broke, or nearly so. They defend this postulate with three arguments. First, Rendall maintains, whereas “rich elderly may accumulate substantial bequests, poorer elderly do not.”

Second, Rendall contends, poverty by definition is involuntary. “If an elderly person chooses to live in poverty in order to leave a large bequest to his children, does the public call this person poor? We argue that you don’t.” Finally, Speare cites current welfare policies. “Most government programs disqualify people if their total assets exceed a very small amount. The assumption is that people should spend all their assets before they ask for assistance,” even though this may leave them destitute for the rest of their lives.

The researchers advocate policies based instead on new types of retirement financing, such as reverse equity mortgages. “You make an agreement with your bank to give up all or part of your home when you die in exchange for current income,” Speare elaborates. Such options are not widely available, however, and the sociologists emphasize that their model is a better measure for the prevention, rather than diagnosis, of poverty. *—W. Wayt Gibbs*




*SOURCE: Michael S. Rendall, Cornell University; Alden Speare, Jr., Brown University; U.S. Bureau of the Census (1984 data).*

LISA BURNETT

this time they are going both up and down. And they think they have a way to break the elevator height barrier by eliminating cable altogether.

Extreme crowding in Japanese cities has drawn attention to the use of underground space. At the Geospace Engineering Center, a subsidiary of the Engineering Advancement Association of Japan (ENAA), companies that include NEC, Fujitsu, Hitachi, Mitsubishi Electric and Fuji Electric have been carrying out preliminary work on these concepts for seven years. ENAA is backed by Japan's powerful Ministry of International Trade and Industry.

The key to the cordless elevator is the same linear motor technology now being applied to the development of magnetically levitated rail transportation systems. Under consideration is a linear synchronous motor that propels a passenger car along the tracklike guideway running the length of the elevator shaft. The motor works by the interaction of an electromagnetic field from electric coils on the guideway with magnets on the car.

In theory, the linear motor elevator, free of cables, can travel up or down any distance at all as long as the power is on. Another advantage might be that the elevator can travel horizontally as easily as it can vertically. In the subterranean city, the same cars that transport urban troglodytes to and from the surface can move them horizontally from home to office or theater or school. Moreover, expensive, space-hogging elevator shafts would no longer be limited to a single car. At any moment, several cars could be descending in the "down" shaft, and as each reaches the bottom of the shaft, it could simply move sideways and then ascend the "up" shaft.

ENAA is also looking skyward in what it calls Tokyo Eco-polis City 1000. This proposed project involves the erection of a building 1,000 meters tall that will house 75,000 people, while providing 700 hectares of office and commercial space. ENAA has proposed the creation of "large-scale man-made structures, rich in greenery and free from the traumas of commuting"—a not-so-oblique reference to the travails of getting home in Tokyo.

Fuji Electric, under contract with ENAA, has constructed a linear motor elevator to test the concept and has recently launched a two-year effort to test a larger model capable of moving on both a vertical and horizontal path. Working prototypes big enough for riders are a long way off. Current plans call for a new model that measures some six meters in height. But that does not stop Japanese engineers from dreaming. Taisei Corporation of Tokyo, one of Japan's most prominent construction companies, has built a small model in its EXCEED 2000 project.

The issue that needs to be resolved is safety. Cut the power to a linear motor elevator and it will drop, though not precipitously. This will require new concepts in braking systems. When brake failure could result in a fall of a kilometer or more, anything less than perfection is unthinkable—and perfection is an engineering parameter that is still in short supply. *—Robert Patton, Tokyo*

## Power Pack

*Batteries are the bottleneck for portable electronics*

Computer and telecommunications manufacturers consider a six-month product cycle an unremarkable phenomenon. For battery makers, eight to 10 years is a better-than-average time to market. This disparity is turning into a barrier for electronic designers trying to condense a panoply of office equipment into devices that can rest within a palm or, at worst, fit the dimensions of an airplane's food tray.

"The battery is holding back additional system improvements," says Michael P. Dooley, director of strategy and business development for Motorola Energy Products. Even Motorola's suppliers do not disagree. William Bowden, a Duracell manager, admits that "the pace of technological change in my industry is glacial."

Packing more power into a mixture of chemicals stuffed into tiny cans labeled AA is a laborious process not conducive to the exponential gains in performance traditional for chip makers. Even so, companies such as Motorola are trying to prod suppliers such as Duracell to try harder. Toward this end, battery makers are attempting to improve their reaction times by focusing on the great light hope: lithium. The lightest metal, it gives up electrons more easily than does any other material used in batteries.

Electrochemists have targeted lithium as the best candidate for providing high-capacity, rechargeable batteries. Although tiny lithium disposable camera batteries are on the shelf in pharmacies, manufacturers of rechargeable batteries have struggled with the tendency of this highly energetic material to smoke or to catch fire when exposed to repeated charges and discharges. This occurs because lithium tends to build up, or "plate," unevenly on the anode (the negative electrode) during the charging cycle.

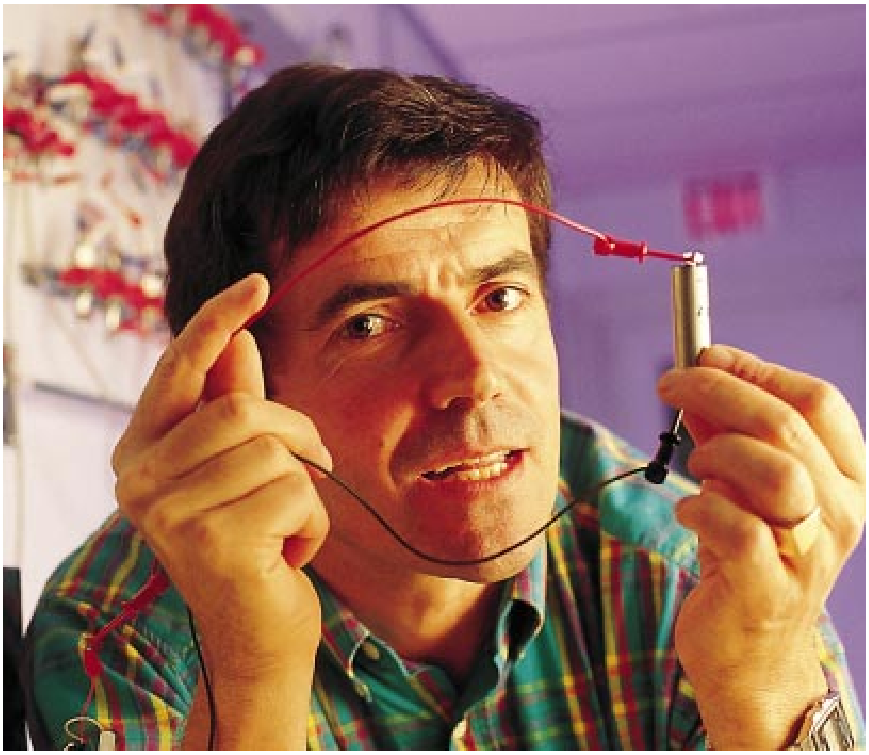
That is apparently what happened when Moli Energy in Vancouver produced more than a million rechargeable cells for Japan's Nippon Telephone and Telegraph. These lithium metal rechargeables, as they are called, were quickly pulled off the market in 1989 after 10 of them "vented"—smoke or flame was released from a battery vent because of a chemical interaction between the plated anode and the liquid electrolyte.

Battery companies have since focused their attention on ways to produce more benign lithium products. One lithium rechargeable that is beginning to reach the market, for example, is called a lithium-ion, or "rocking chair," battery. During the charging cycle, a lithium ion "rocks" between a cathode (positive electrode) made of a material such as an oxide of cobalt, nickel or manganese and an anode of carbon or graphite. The important thing is that lithium-ion cells do not experience the dangerous buildup of metallic lithium on the anode. Rather, the lithium atoms insert themselves between layers of graphite or carbon in the anode, a process called intercalation.

These lithium-ion cells store two to three times the energy per weight of conventional nickel-cadmium cells. The voltage—3.6 volts—enables a portable-electronics designer to gain a much-sought weight saving for a computer or camcorder because one cell can be used instead of two or three. But circuitry used to recharge nickel-cadmium or nickel metal-hydride cells needs to be redesigned for the lithium cells.

Development of the lithium-ion battery has become the preoccupation chiefly of Japanese electronics manufacturers. Their domination of this technology, like the lead they have snared in flat-panel displays, ensures close compatibility between component engineers, who draw up plans for individual batteries, and systems engineers, who design the consumer products powered by these cells. "We're not only behind in the manufacturing but in the basic research for these batteries," says Jeffery R. Dahn, a professor of physics at Simon Fraser University, who was formerly research director for Moli Energy.

Evidence of the U.S. lag in development could be seen at the meeting of the Electrochemical Society in Honolulu in mid-May. Sanyo and Matsushita reported increases as high as 50 percent in energy-storage capacity for lithium-ion prototypes over a commercial version of the battery that Sony Energytec, a subsidiary of the electronics giant, has begun incorporating into camcorders and cellular phones. Besides Sony, A&T Battery (a joint venture of



ROBERT PROCHNOW

***GREAT LIGHT HOPE is researcher Jean-Marie Tarascon's ambition for this lithium-ion cell that he helped to develop at Bell Communications Research.***

Asahi and Toshiba) was slated to scale up to production of hundreds of thousands of cells a month.

The only U.S. institution to present a lithium-ion paper at the conference was Bell Communications Research (Bellcore), the research arm of the regional telephone companies. Bellcore is interested in a battery that could serve as a backup power source for the optical-to-electronic conversion devices for fiber-optic communications. Jean-Marie Tarascon, leader of a Bellcore team, fashioned a cathode from lithium, manganese and oxygen. Manganese is less expensive than the cobalt in the batteries developed by Japanese manufacturers. Bellcore, which like its Baby Bell parents is prohibited from manufacturing, is licensing the technology to an undisclosed U.S. company. The big three U.S. manufacturers—Duracell, Eveready and Rayovac—also have in-house research teams working on lithium-ion batteries.

The Japanese lead in lithium-ion batteries may cause U.S. companies to concentrate on another lithium technology still secreted in the laboratory. This material is a lithium-polymer combination. Anodes of lithium-polymer batteries consist of a layer of thin lithium foil alternating with a solid polymer electrolyte, such as polyethylene oxide. A material like vanadium oxide is used as the cathode.

If substantial developmental barriers can be overcome, engineers crafting the new lithium cell dream of placing the thin-layered cells in shapes that allow them to conform to the design of a phone or computer case. Amalgamating these films into a producible battery still remains an unmet challenge. But the industry hopes that machinery for making cellophane and other plastic films can be adapted for the high-speed manufacturing of polymer batteries.

A lithium-polymer battery could exceed the energy capacity of its lithium-ion cousin because it need not carry the burden of a carbon or graphite anode that acts merely as a storage material for the lithium. Instead lithium would constitute the entire anode. But such a solid-state battery will become a reality only if engineers can find polymers that are more conductive and that do not degrade after 150 or so cycles.

Some lithium-polymer developers have tried to make the batteries conduct more electricity by adding small amounts of a conductive liquid solvent to the polymer. But the cell becomes a candidate for the same safety problems experienced by Moli Energy.

Basic research on the polymer electrolytes is inching forward. A team at Arizona State University found that adding a small amount of polymer to lithium salts produced conductivity comparable to that found in many liquid electrolytes. But the rubbery "polymer-in-salt" electrolyte is still too reactive, and it tends to crystallize. Despite the uncertain prospects, a Long Island company, Moltech, has already agreed to take out a license on the Arizona State electrolyte. 3M and others have also shown an interest in negotiating a license.

The risks in new battery development have led to a bumper crop of consortia. The consumer market could, in fact, receive a much-needed boost from the U.S. Advanced Battery Consortium (USABC), a joint effort of automakers and battery manufacturers to develop lithium-polymer and other batteries for an electric vehicle. In addition, the Microelectronics and Computer Technology Corporation in Austin, Tex., has been trying to gather battery and electronics manufacturers into a collaboration that would adapt lithium-polymer technology to portable devices. And the Advanced Research Projects Agency plans to distribute \$5.7 million largely for research on lithium-polymer cells.

Getting battery companies, consumer electronics manufacturers and national laboratories to work together has often turned into a legal morass. It has taken Sandia National Laboratories more than two years to set up a three-year consortium, with \$300,000 in annual funding, to adapt a carbon material created for the moribund "Star Wars" program into an electrode for lithium-ion batteries. Duracell, a prospective participant, was unable to reach agreement with Sandia and so was left out of the collaboration. "The unfortunate thing is that when we started, we had a material that was two times better than what Japan had, but during the period of negotiating with the lawyers, the Japanese have made a lot of advances," says Samuel C. Levy, a member of the technical staff at Sandia.

An upstart in the lithium-polymer business is Valence Technology of San Jose, Calif. A securities analysis report by a Valence stock underwriter raised eyebrows by suggesting that the batteries could be produced for less than a dime for each watt-hour of energy. "That's less than what they would pay to buy the lithium," Dahn comments.

Even so, Motorola signed an exclusive agreement with Valence to buy \$100-million worth of the start-up company's batteries if Valence can meet the big electronics manufacturer's technical and cost specifications. The announcement sent the company's stock soaring for a time. "We have nothing committed up front," says Dooley of Motorola. "The message we're trying to convey is that it's pretty clear that Motorola doesn't want to sit around and wait until the battery industry does something new." *—Gary Stix*

## A Digital Fix for the Third World?

The notion that technology can fix the economic problems of the developing world dates back to the early years of the post-colonial era. The "green revolution" of the 1960s held out the hope that the labors of Western plant geneticists could eliminate food shortages by raising crop yields.

The information age has begun to raise similar expectations. African heads of state, such as Flight Lieutenant Jerry J. Rawlings, who staged the overthrow of a civilian government in Ghana more than a decade ago and last year was elected as president, sound a little like executives from Silicon Valley. Rawlings now talks about an "enabling environment" for efficient telecommunications in a country where many employees spend about half their time going from one place to another because of the unreliability of the telephones.

Rawlings's neighbor, Felix Houphouet-Boigny, longtime leader of Ivory Coast, voices a similar hopeful refrain: "Thanks to informatics, technological shortcuts to development exist today and are within the reach of everyone. We are not therefore doomed to remain undeveloped forever."

For all its promise, the allure of skipping over several generations of technology has fallen into disfavor in some circles. Economists often adhere to the "product cycle" approach to diffusion of technology. This trickle-down theory holds that developing nations absorb new technology only after it has become widely established in industrialized countries. Costa Rica assembles baseballs for the major leagues; Du Pont turns out aerospace composites. Other economists also point to the billions of dollars wasted on aid that recipients have been unable to integrate into their economies, from tractors without gas stations to steel mills without transport for ore or finished products.

Many preach an enlightened cousin of the trickle-down approach known as appropriate technology. And the information advocates argue that high-tech electronic tools are appropriate—$1,000 microcomputers can do the work that required $1-million mainframes in the industrialized world only a decade ago, and cellular telephones can fend off the lengthy process of laying land lines.

Telecommunications linkages in Ghana, for example, might tie cocoa producers more closely to customers in world markets, allowing them to wheel and deal more rapidly and to increase profits that could eventually flow back to farmers. Service jobs that require manual processing of data are being moved from back offices in high-priced countries to provide jobs for city dwellers in the developing world.

*Information technology may bring limited benefits to the work force of developing nations.*

Moreover, networks could keep academics in developing nations in touch with colleagues in the North. The National Science Foundation is contemplating the funding of U.S. university programs to set up data banks that would let both U.S. and Latin American engineers and scientists share information over the Internet data network.

Dogging all these new-age ideas is the fact that the green revolution and other technological salvations were not exactly unmitigated successes. Agricultural aid definitely improved crop yields for larger farmers with the know-how and capital to buy and use hybrid seed, fertilizer and the other accoutrements of a midwestern U.S. farm. But in India and Pakistan, the poorest subsistence farmers and unlanded laborers were still left on the margins.

Similarly, the benefits of PCs and pocket phones have been confined to a narrow segment of society in developing countries. In India, a more than $200-million software industry has emerged, centered around Bombay and the southern city of Bangalore. Software houses there have crafted, among other applications, programs that let U.S. consumers buy gasoline automatically with a credit card, according to Pyramid Research, a consulting firm in Cambridge, Mass. Yet less than than 1 percent of Indians have phones, compared with 96 percent of Americans. Most branches of the country's largest bank, the State Bank of India, still record transactions by hand.

"In these places, what you may see are enclaves of high-expertise engineers and high-skilled technicians with telecommunications links to wherever they're getting their orders—and a barbed-wire fence between them and the rest of the economy," says W. Brian Arthur, a professor of economics at Stanford University. "The proceeds of what they do may go to some Swiss bank, so they may as well be on the moon."

Africa also has measured some gains from information technology and telecommunications, but a huge gap remains in the ability to meet most countries' needs. The cocoa board in Ghana uses Morse code to communicate with offices all over the country.

The government in Ghana now allows private companies to compete with its own Posts and Telecommunications Corporation, which provides 48,000 telephone lines, according to an article in a journal called *African Technology Forum*. Ghana could accommodate at least 200,000 telephone lines, observers estimate. But at $1,000 to install a handset—more than the nation's average per capita income—demand for privately supplied telephones will remain limited.

Instead, where information technology has become a boon is in the newly industrialized nations, those with the organizational and planning skills to make full use of the technology. Singapore saw a dramatic increase in revenues from software and services during the 1980s. The small nation's advances resulted from an aggressive government policy of building telecommunications networks and promoting a domestic software industry. Other nations—Venezuela and Mexico are two—have witnessed a rise in supply of telecommunications services after privatization of government telephone monopolies.

Indeed, some economists worry that information technology may actually draw jobs away from the developing world, further exacerbating the gap between rich and poor. Just-in-time inventory control methods and computerized manufacturing have become a form of intellectual and organizational capital that outstrips the importance of raw labor costs. In short, Ashoka Mody, a World Bank economist, warns, "the low wages in the developing countries can get lost in the noise of cost accounting." —*Gary Stix and Paul Wallich*

# Making Fluids into Solids with Magnets

Anyone can turn a liquid into a solid using an ice tray and a freezer. But can you solidify a fluid and then liquefy it within a few seconds? Actually, all you need is a strong magnetic field and a simple recipe for a magnetorheological fluid.

Such mixtures consist of particles suspended in a fluid. When the substance is exposed to a magnetic field, the particles align in such a way that the mixture becomes solid. When the field is removed, the particles return to their random state, and the substance liquefies. The degree of solidification depends on the inherent properties of the fluid and the strength of the applied magnetic field.

Magnetorheological fluids are not as common as a similar class of materials that solidify in the presence of electric fields. These electrorheological fluids have been incorporated in prototype brake systems, clutches, shock absorbers, engine mounts, actuators, control valves and even artificial joints [see “Electrorheological Fluids,” by Thomas C. Halsey and James E. Martin, page 58].

DANIEL J. KLINGENBERG is an assistant professor of chemical engineering at the University of Wisconsin–Madison. He received his Ph.D. from the University of Illinois at Urbana-Champaign and has written several papers on electrorheological fluids.

Yet amateur scientists would be wise to experiment with magnetorheological fluids before they attempt anything with the electrical counterparts. Most electrorheological fluids require high voltages. For example, several thousand volts would be needed to solidify a mixture of corn oil and cornstarch. (The actual amount of electric current that would flow through the liquid would be quite low, however.)

Although present magnetorheological fluids are not as versatile as their electrical cousins, they do demonstrate the effect that so excites many researchers. The experiment described here can be performed on the kitchen table for less than $20.

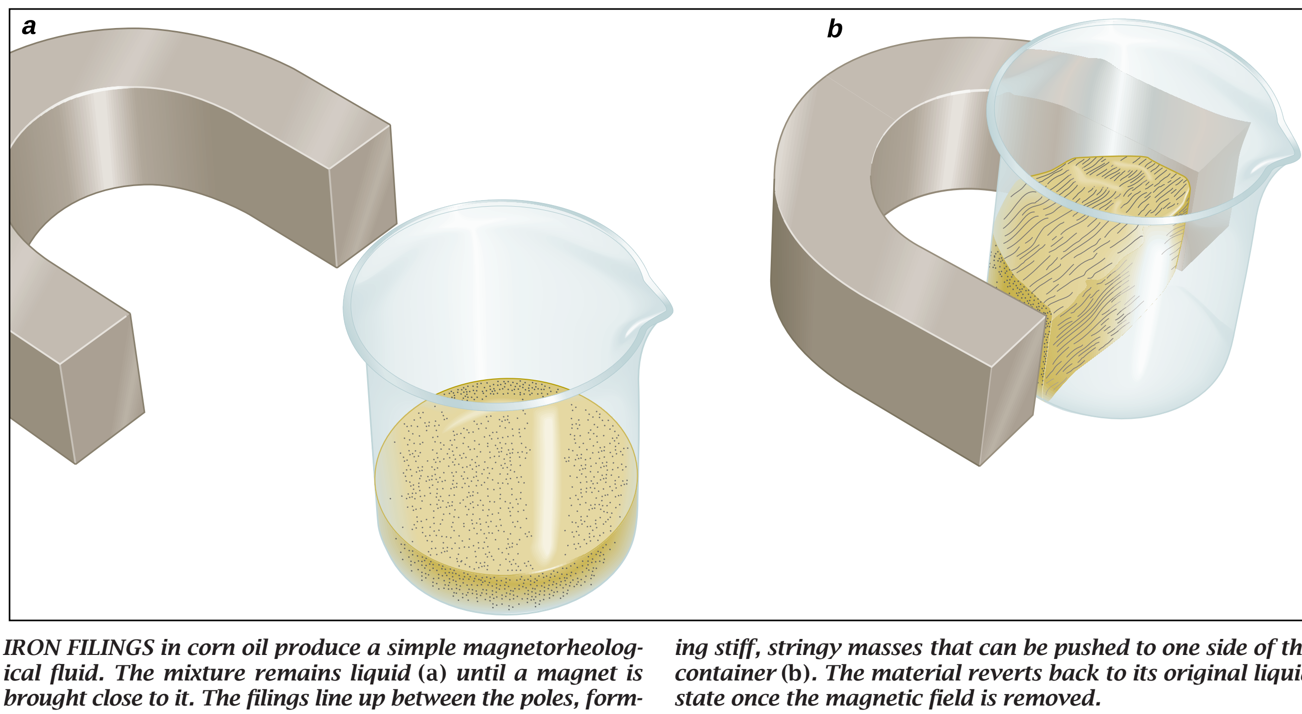
A simple magnetorheological fluid is made of iron filings and corn oil. You can find iron filings at a toy store, at your workbench or at a machine shop. Specifically, you could empty the filings out of a toy magnetic sketch pad; you could scrape a piece of iron with a good metal file; or you could gather the iron filings produced by a metal lathe.

The size of the filings will determine the magnitude of the magnetorheological effect observed. The filings should be large enough to identify single particles with a magnifying glass but preferably less than half a millimeter long. Filings that are too large can easily be ground to a smaller size with a spoon and a hard surface. Be aware that iron will scratch the surface.

The type of oil needed to suspend the filings does not matter much. It just has to be more viscous than water, to slow down the rate at which the filings settle. I have used corn oil, linseed oil, silicone oil and motor oil. Corn oil is perhaps best: it is cheap and cleans up relatively easily with soap and water.

The filings and oil should be mixed in a small, clear container, such as a clean aspirin bottle, a plastic cup or, as I used, a 100-milliliter beaker. Just make sure that you can see down into the bottle. The fluid should consist of approximately 25 parts by weight oil to 100 parts by weight iron filings. The exact proportions of the constituents, however, are not critical to observing some kind of magnetorheological effect.

As for the magnetic field, almost any type of strong, permanent magnet will

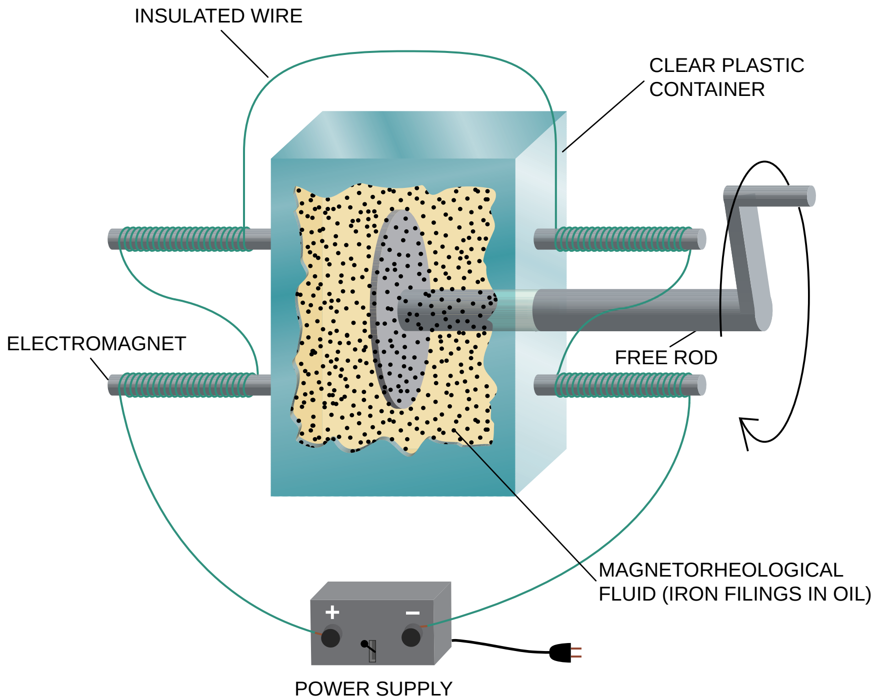



***IRON FILINGS in corn oil produce a simple magnetorheological fluid. The mixture remains liquid* (a) *until a magnet is brought close to it. The filings line up between the poles, forming stiff, stringy masses that can be pushed to one side of the container* (b)*. The material reverts back to its original liquid state once the magnetic field is removed.***

work: horseshoes, bars, even the ones from old loudspeakers. (Refrigerator magnets are typically far too weak.) Magnets are often rated by the number of pounds they can pull. You should buy the most powerful ones available. I purchased two "100 pound" bar magnets from a local hardware store for about $12 apiece.

Electromagnets are also an option. Such magnets are made by winding thin, insulated wire around a metal bar and connecting the ends of the wire to a battery. The magnetic field will be strongest near the ends of the coiled wire. It does not matter whether you wind the wire from one end to the other or back and forth several times. Just be sure that the wire is always wound in the same direction (that is, clockwise or counterclockwise) and that the winding is tight. Two layers of windings over a couple of inches of the core should suffice.

To amplify the strength of the magnetic field, you can increase the power supplied to the electromagnet by connecting several batteries in series. Note, however, that the electromagnet draws a lot of current and will quickly drain the batteries. A power supply that can deliver several volts may be a better choice. Although it is quite fun trying to create a good electromagnet, it is hard to beat a permanent magnet in terms of both time and convenience if your goal is to play with a magnetorheological fluid.




***MAGNETORHEOLOGICAL BRAKE relies on a disk mounted on a rod. Turning on the magnetic field prevents the rod from rotating. Batteries rather than a plug-in power supply can be used, although they will be drained quickly.***

Once you have your fluid and magnet ready, you can observe magnetorheology. First, stir the fluid with a piece of wood or plastic (the filings will adhere to a magnetizable object), keeping the magnet at least six inches from the container. You should notice that the combination of filings and oil behaves as a fluid: it continues to fill the container and tends to form a smooth, flat surface after you stop stirring. It has the consistency of maple syrup.

Bring the magnet near the container. If you are using a horseshoe-shaped magnet, put the container just inside the two legs. If you have two magnets, place the opposite poles on either side of the container, facing each other. The closer the poles are, the stronger the magnetic field in the fluid will be. If you have just one magnet, place one of the poles directly on the side of the container.

Stir the fluid again and notice how it begins to behave like wet sand. Distinct, stringy clumps should form at the bottom of the container. These clumps should retain their shape and remain suspended in the fluid after you remove the stirrer. The clumps appear stringy because the filings tend to attach end to end to one another, extending from the poles of the magnet. Now move the container away from the magnet. The filings-and-oil mixture will rapidly change back into a fluid: the clumps fall apart, and the surface becomes fairly smooth and horizontal. If you do not observe the effects, you probably either need to get a stronger magnet or need to change the consistency of the fluid.

As a variation, repeat this experiment with different proportions of the filings and oil. The optimum combination will depend on the size of the filings and the type of oil. Also, attempt to vary the distance between the poles of the magnet and the container or else use a stronger magnet.

A much more challenging project is to create a simple magnetorheological brake [*see illustration above*]. Machining skills may be needed here. Insert the end of a rod into a plastic container filled with the magnetorheological fluid. The rod should be made of something that cannot be magnetized, such as plastic or aluminum. You might try attaching a plastic disk to the end of the rod in the fluid, because a larger surface area in the fluid will yield a more pronounced effect. To keep the fluid from leaking out, you may need to place a rubber O-ring around the rod where it enters the container.

Now all you need to do is immerse the container in a magnetic field. If you use electromagnets, I suggest wiring several perpendicular to the disk in the fluid. Be sure to wind the wire around each core in the same direction.

Before the magnetic field is applied, the rod rotates with little or no resistance. (To make the rotation more visible, you may wish to bend the end of the rod or attach some kind of handle.) Applying the magnetic field solidifies the fluid. The rod is now difficult to turn.

As you may notice with this combination of iron filings and corn oil, such fluids often do not solidify robustly and tend to break down over time. Research in industry and academia has made significant progress in improving these fluids for applications. For example, the Lord Corporation, based in Cary, N.C., plans to market magnetorheological fluids that promise to use much less magnetic energy to function than the formulation described here. But several years will pass before these fluids will drive and stop our cars.


FURTHER READING

FERROHYDRODYNAMICS. R. E. Rosensweig. Cambridge University Press, 1985.

ELECTRORHEOLOGICAL FLUIDS: PROCEEDINGS OF THE SECOND INTERNATIONAL CONFERENCE. Edited by J. D. Carlson, A. F. Sprecher and H. Conrad. Technomic Publishing, Lancaster, Pa., 1990.

## The Demographic Transition

**A Concise History of World Population,** by Massimo Livi-Bacci. Translated by Carl Ipsen. Blackwell Publishers, 1992 ($49.95; paperbound, $17.95).

This readable account by a Florentine demographer and statistician is plainly in the highest civic spirit. Massimo Livi-Bacci displays the main strands of the thick skein of human head counts unusually well.

The book opens with the simpler mathematics of the discipline, expounded by example. Then it puts side by side the conflicting explanations that have governed attitudes toward population growth ever since the days of Reverend Thomas Malthus. Fully half the text treats a striking contrast that has opened between countries rich and poor and its decisive correlates, both social and economic. Throughout, the author grounds his views on the numbers, well chosen and meaningful numbers that he spreads before us in 60 diagrams and figures and a couple of dozen tables.

Jump first to news exciting enough to warrant public celebration in the streets. Across one graph field, two curves compare two groups of countries, the rich with the poor. The curves plot the smoothed annual percentage rate of population increase over the past three centuries. We all know that the industrialized countries have managed control over both birth and death. Their net growth curve peaked broadly around 1900 at a rate a little more than 1 percent a year and has by now fallen about fourfold. *The poor countries too have entered the same transition.* Their peak percentage rate of growth was reached as recently as 1970 and has fallen steadily ever since, much more steeply than the comparison peak of 1900 among the well off. To be sure, it had risen to almost 2.5 percent a year, so it has a long way to fall.

Each major region of the developing world taken separately also shows a clearly declining growth rate, save only Africa, where the growth rate is currently about at plateau. The author says firmly: "The fact remains that demographic change in the poor world in recent decades has on average proceeded rapidly as compared to the path previously followed by the rich." Of course, no extrapolation can itself give a guarantee, but all signs are strong and clear. Absolute numbers will grow as the abounding young become parents, but relative rates have begun a steady decline.

For demographers, our common future is played out across an abstract "strategic space," a simple if fateful plot whose vertical axis shows the average number of children born per woman, as its horizontal one marks the average expectancy of life at birth. How these are estimated is summarized. The two parameters are far from independent, for mortality among young women of course reduces the number of children expected. (Note that abundant sperm matters here much less than rarer eggs; in less biological discourse the societal status of women is a major index to demographic realities.) On such plots, concave curves that dip from upper left to lower right mark the locus of all points that correspond to a particular annual percentage growth. Historical populations from the Paleolithic (largely conjectural), firmer numbers from 18th-century Europe, on to India and Russia about 1900, and a dozen more dot a crescent-shaped band with annual net growth rates all between 0 and 1 percent. That is the old enduring regime, once universal.

In contrast, another plot shows the strategic shifts that have occurred since 1800 among 17 countries in Europe. The old ellipse that then enclosed all their dots has slid steadily downward and rightward. By 1980 they all occupied a still tighter ellipse, centered on a small negative growth rate. Once the number of children per woman clustered between four and seven, and the expectancy of life held around 30 years. The 1980 values show an average of fewer than two children per woman, and an expectancy of life beyond 75 years. City life dominates now; the demand for children's labor drops; the cost and length of their education rise; women have more independence and

**CLIFF DWELLERS,** ***a 1913 painting by George Wesley Bellows, is used as the cover illustration of*** **A Concise History of World Population.**

# 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.


marry later in life; the control of births rivals the postponement of death. Survivorship has "increased immensely."

A good century ago only one 15-year-old in 10 could expect to reach the age of 80; now about half of us do. But most of the aging of the population has come not from better survival of the old but from reduction in the young population because of so few recent births. From now on, improvements in mortality will need to benefit the old; there is "no room for improvement" among the young, for 98 percent of those born reach reproductive age. (These are figures for Italy; the U.S. with its cruel legacy of slavery still has ample room to better the vital statistics of our children.)

The tribute to war, famine and pestilence declines, too. A graph shows the crude death rate for Sweden between 1740 and 1920. Before 1820 we see three deadly peaks, each a few years wide, during which death rates more than doubled. As the decades passed, all the peaks became pinched between two trends—a slowly falling baseline and a faster falling maximum of deaths. There is by 1920 a 10-fold decline in the variation of the death rate. The new regime has begun, substantially relieved from the fear of sudden death. Grim war remains as do emerging diseases, but as long as we fight off novel disease and keep the nuclear peace, their episodic toll will be small.

Between 1950 and 1985 the dots that describe two dozen of the most populous of the poorer lands show a general procession along the well-marked European route, not yet nearly as far along as are Europe and the neo-Europes overseas (plus Japan). A worldwide demographic transition is patently in full progress. Conceding at least one doubling of the head count in the next century, we still don't face runaway growth.

What about theories? In these complex human affairs, so rich in feedback loops and in feed-forward anticipations, it is no easy matter to tease out true causes. Old Malthus could argue logically that population growth with diminishing returns to farmers' effort would reduce per capita production, increase poverty and bring both hard times and a population decline. So it can; indeed, the scheme fits the cycles of his time in France and England. But about a generation ago, another scholar of agricultural economics, Ester Boserup, argued quite differently and equally convincingly. He held that the demand for more food calls out more intensive labor and new production methods, and hence, eventually, more food, and then still more people at the table. "This dilemma is unresolvable only if we insist on finding hard and fast rules to explain complex phenomena." Both outcomes are only contingent scenarios, not fateful compulsions.

Real human adaptation, studied here over a variety of economic and social variables, is much more subtle. Our author sees the chronicle as a process of choices and constraints, both extended over time, neither of them ruled by any single simple factor. Hard constraints are surely real; more than a billion of the poorest of us humans still live hoe in hand under the old regime. And choices are clearer; in China by mandate and in easier India as well, limitation of fertility is genuine. All save a few of the national dots drift predictably along the decisive path to population stability. The bottom line a century ahead is not easy to calculate, but the wide trend to workable control is plain, a trend not only in our growing abilities but in the increasingly common purpose of the improvement of human life.

## 16th-Century Shipwrights

**VANGUARD OF EMPIRE: SHIPS OF EXPLORATION IN THE AGE OF COLUMBUS,** by Roger C. Smith. Oxford University Press, 1993 ($35).

Portuguese ports have a resource unique in Europe: by departing on the summer trades, ships could sail downwind toward the ocean islands and America and yet count on fair winds out of the west to speed them home again as the seasons passed.

Young Prince Henry the Navigator soon determined to press his country's future at sea steadily southward along the African coast. He died in 1460, about when his skilled captains had first felt their way to the equator, but his pioneering quest survived him. Portugal and her neighbor, Spain, alone would send their flags around the world, until by about 1600 their empires experienced overt competition from the other states of Atlantic Europe. No small part of Portuguese seapower came from that early understanding of Atlantic winds and currents.

These Iberians, major early users of inanimate energy, moved by windpower and fought by gunpowder. In this crisply focused narrative we are shown in some detail how it was all done, from the loans made by Italian bankers to the master lodestones kept in Seville to prepare strong magnetic needles for the pilots. We read chapter by chapter just how the ships of Lisbon and Seville were built, then rigged and outfitted, manned, provisioned and, of course, armed. These self-assured and ruthless explorers were paramilitary; their superior weapons gave them victory, to the east against all marine powers from the Swahili ports to Japan, to the west against no real opposition at sea.

The details are rich. But the freshest topic any landlubber will find here is how we come to know all this and how well we know it now. We have shipyard contracts and legal inventories, copious sealed state records of all the voyages and what they found, artist-embellished maps and texts that bear plenty of images of the outsides of ships. But we have little counterpart to blueprints from those wonderful craftsmen whose skills, tools, sketches and rules of thumb fashioned the complex hardware of planetary enterprise. After 1600, scholars began to describe all they could learn. Modern Iberian scholarship in particular has sifted the records, recalculated ship performances, compiled technical glossaries and sought the sources and nature of wrought-iron cannon, hempen canvas and saltpeter, even to Columbus's own experienced recommendation for "the third part of the breadstuff to be good biscuit...and not old."

What is most wonderful is that within the past decades we have found even more intimate and reliable data not on paper but in this world of three dimensions. A meticulous ship's model was given in votive offering long ago to a church near Barcelona. The four-foot carved model represents an actual medium-sized ship, probably a little earlier than those of Columbus. The model, now in Rotterdam, has been studied inside and out by medical optical instruments, videotaped and itself replicated. It solves several knotty puzzles of ship construction.

A still better time machine is the old shipwrecks found on both sides of the Atlantic [see "Raising the *Vasa*," by Lars-Åke Kvarning, page 84]. A few vessels of about the period have been unearthed along the Po and in England. One proud battleship of Henry VIII, a royal carrack, capsized and sank in 1545 while readying for battle with the French off Portsmouth. It was resurrected a decade ago. The coasts of the New World have turned up remains that are even closer to the issue. On the Yucatán coast, in southern Texas, then north to the Caicos reefs and the Great Bahama Bank, the divers have found and the scholars identified half a dozen merchantmen of the Spanish Main, all from the first half of the 16th century. On the sea bottom, they look like oval mounds of boulders, for a pile of waterworn stones ballasted every ship of the day.

Around, above and beneath those stones lie the hardware, weapons, stores and whatever is left of the wooden ship itself. One Basque venture into whaling sank in the cold waters of Red Bay, Labrador, after slipping its moorings late in 1565: a rubber casting of the center of the ship now preserves full-sized a fine example of "sixteenth-century shipwrightery." An exciting summation ends this engaging book.

But the handy fore-and-aft sail of the caravels, the sternpost rudder, the magnetic compass, the cannon, gunpowder and crossbows that armed these ships for war are all of them inventions of Chinese origin, practical for centuries before they found the way west. The author, perhaps too intent, has not recognized that fact by a single word, even as an aside. Not invention but development is here on center stage.

## A Swift Viral Traffic

EMERGING VIRUSES, edited by Stephen S. Morse. Oxford University Press, 1993 ($39.95).

A thousand centuries ago our sapient forebears first quit the ancestral home, leaving a life pattern wherein protohumans had long equilibrated with local infections. They multiplied greatly in the temperate world—Eden until Cain, prolific farmer and stockman, would largely dispossess his foraging sister and hunter brother. The old disorders of our scattered kind turned into a fitful tale of universal childhood infections in the more populated settlements. Those pathogens survive on numerous new human births, even though their attacks confer lifetime immunity on each survivor: witness measles and smallpox. In the past century another grand flourish of human numbers appeared, diseases kept at bay by a thin modernity, whether in medicine, in food or in well water. Today we restlessly transcend all geography at jet speed, while farmers still sow and reap many new fields at the feather edges of living forests. Of course, a swift viral traffic challenges what we once imagined as a balanced and tolerant regime.

It is influenza A that causes the teary, achy fever a hundred million of us undergo every year worldwide. Its agent is an old but steadily renascent virus, surely the respiratory epidemic clearly described in Europe since medieval times. Out of the east it came again and again, to rage and to go away in a matter of weeks or months, and still it does. But now we understand its RNA core and its surface proteins, tiny spikes that induce the human immune system to form antibodies. The antibody record in the blood of elderly patients demonstrates that since 1890 all influenza epidemics have come from only a few of the possible subtypes of the virus particles (we do not know why).

The virus genome can mutate, changing the molecular code that forms the spike-tip proteins, so that the virus is able to evade immune attack until new custom antibody is formed. But the principal changes we see year after year in epidemic human flu are far less subtle; they result not from merely defensive point mutations at the spike ends but from the analogue of chromosome assortment among the eight gene segments of the viral core of RNA. A new mix of genes can bring not simply a changed infectivity but new virulence. With a modified protein the particles may no longer cause a self-limited illness but can instead turn more deadly by one or even two powers of 10.

We share influenza A with our domesticates—pigs, turkeys, chickens and horses. The infection can move both ways; all U.S. hog farms have sick pigs each fall. Sometimes their virus shares the spike type of the human strain-of-the-year, although swine virus usually lacks the ability to spread among humans. Whence, then, the new flu almost every year?

"The ultimate reservoir of influenza is in the aquatic birds of the world." All 13 or 14 subtypes of influenza are found in wild ducks, where they cause no disease at all but replicate temperately, mostly in the duck intestinal tract, to spread and mix within hospitable waters. "Every August you can go to Canada and...isolate your own kind of influenza virus." Wanderers like gulls and turnstones maintain the flu virus in the spring; migrant ducks harbor it in the fall. An epidemic of flu among New England harbor seals in 1980 clearly derived from that avian reservoir of virus. (The severe seal plague of 1988 in Europe was not the flu but a virus related to canine distemper and to the terrible cattle rinderpest.)

Airborne transoceanic spread of flu virus did not wait for Boeing; avian virus types suggest two ancient flight paths. The path is north-south from Siberia in the Old World and from the Canadian Arctic southward in the New; only a little mixing occurs across the Alaskan Straits. The inference from long field studies is now strong: the gene reassortment that year after year generates new types of human flu proceeds in those myriad ponds where the farmers of Asia grow paddy rice. Within those unfreezing waters, ducks, pigs and humans together shed and acquire virus particles.

Passage through avian host cells usually selects for virus genes that do not spread well between humans. But the potential for a murderous pandemic is always present in those chancy, infectious and indispensable ponds. The influenza A pandemic of 1918–19, called in error "Spanish" flu, turned out to be both highly infectious and uniquely virulent. That edition of the flu took as many human lives as any single war has ever done.

This volume presents 28 brief chapters in dazzling variety, only a handful of them in too narrow a jargon. Among the authors you find a historian, a mathematician, clinicians, virologists, evolutionists, a student of mosquitoes, epidemiologists with strange tales from the Pampas or the West African bush, molecular biologists and a few very worried senior public health officials.

Troubling lists of a dozen newly emergent viruses are here. As an instance, dengue fever is making a reappearance in the Americas under new urban conditions. Used tires arrive in Texas from Asia by sea; some are piled up outdoors, soon to house tiny pools of rainwater, each perfect to awaken the stowaway eggs of a proved vector of dengue fever, the Asian tiger mosquito. A science-fiction episode is presented here, an exciting, serious extrapolation of what a virulent new tropical disease might mean to the U.S. Our country is slowly becoming unready; our tropical expertise and our labs for virulent infections are for the most part the veterans of old wars. Nor does the world have an adequate virus early-warning system, a clear international need that would double the scope of the World Health Organization, an agency whose total budget is now about the size of that of a single big urban hospital.

Viral RNA lacks proofreading and editing adaptations common in the longer, stabler DNA genomes. That ensures many replicative errors, in analogy to the earliest days of still unstable life. Viruses recklessly gamble in a risky game we more elaborated creatures cannot afford to enter. We still note frequent interaction between cell DNA and virus RNA. Such transfers mark our entire evolutionary history; we dare not now underestimate what viral change can yet do. Equilibrium is no longer seen as unchanging refuge; the peaks and troughs of life's endless novelty demand vigilance and insight.

ESSAY by Richard Wassersug

# Tadpoles from Heaven

Like the cat that came back, the space station *Freedom* once again grips the attention of the U.S. Congress. Year after year Congress tries to snuff out the space station, and year after year it squeaks through. What could be more surreal than a space station that doesn't exist yet won't go away?

Why won't it? Given the more pressing problems on the home planet, why is such an adventure contemplated for even a nanosecond? The human exploration of space will be (excuse the pun) astronomically expensive. We have all heard that robots could do the job more cheaply and more safely. Why, then, are we, the taxpayers, repetitively entertaining multibillion-dollar initiatives to get a few of our kind into orbit, off to the moon and eventually to Mars? Why have we not declared these latter propositions positive lunacy (in the metaphorical sense) and left them at that?

The above are not just rhetorical questions—they deserve answers. I feel some responsibility for finding those answers since I happen to be involved with research ultimately related to human interplanetary exploration. Last year I collaborated with scientists at the National Aeronautics and Space Administration on a project to raise tadpoles in zero gravity. The mission was a success. With the help of the astronauts on board the space shuttle *Endeavor,* my colleagues at NASA produced the first vertebrates conceived in space and then brought to the earth as free-living organisms. My part of the project was to determine whether those little aliens, "our tadpoles from heaven" as I took to calling them, were normal. Basically, they were.

Thus, we now know that it is possible for a lower vertebrate to be born in space and, if brought to earth while still young, to develop normally. It remains to be determined whether mammals can produce normal offspring in zero gravity. But the results to date with amphibians are promising. At least, they do not yet preclude peoplekind or other kinds of organisms from residing beyond the earth's gravitation and leading normal reproductive lives.

The core question behind my NASA collaboration is whether gravity is essential for the development of vertebrates. I consider this to be a serious topic worthy of investigation, whether or not Congress ever approves a moon/Mars initiative. As an earthbound biologist, who began his career as a naturalist, however, I have typically viewed NASA's activities with cynicism. With so much to learn about the universe, I could not understand why NASA undertook what so often seemed to resemble pure publicity stunts.

I remember, for example, learning, back in 1969, that an astronaut had just swung a golf club at a golf ball on the moon. My first thought was, "What a dumb use of taxpayers' money. The guy's wasting millions of dollars when he could be doing serious science. Fire him! Let him find his own transportation home." My involvement with NASA and space biology has not much altered my skepticism about the agency's commitment to science. But it has forced me to ponder NASA's relationship to the populace at large.

What the space agency has provided us with is a primal link to the gods. It isn't science, but it may still be worth millions. Think about what humankind has invariably asked of religion and what we now get from NASA: we speak of heaven "above," with the earth below. Through religion, we have historically "transcended" our earthly burden. The gods were in the skies, and the names of the planets still remind us of that. Before we conquered gravity with rockets, religion was our only path of ascent above the mundane.

Today we treat our shuttle flight crew with the deference that was reserved for high priests in previous centuries. Astronauts are now the intermediaries between ourselves and something still intangible, but unquestionably higher. Our astronauts, like the priests of old, will visit Mars. It is Mars the planet, not the god, but the emotional appeal is still the same.

We have to be careful not to deify our astronauts completely as they ascend to the heavens. Golf on the moon may be dumb, but it is a uniquely human activity played explicitly on the surface of the planet that, good, bad or otherwise, defines our place in the universe. When the astronauts who travel into the heavens also do such human things, they reinforce the humanity that links us to them and thus to the heavens above.

It is not totally clear what kind of scientific information we will gain from a visit to Mars, but I suspect that the trip will take place in my lifetime. It will probably take a quarter of a century and more money than anything that humankind has attempted before, but it will happen. It will happen not because it is essential for our survival and certainly not because it will answer any pressing scientific questions. It will happen because it is a step toward the gods. In sum, space exploration elevates our existence, and NASA sanctifies our lives in a way that no other government activity could ever hope to do.

Because of its high cost, however, any mission to Mars will probably have to be an international endeavor. It could even be argued that to start planning this *hajj* to the heavens now could be a positive step toward global cooperation. Dare I say that one small step for NASA could become, so to speak, one giant leap toward peace on earth and goodwill toward humankind?

Despite all that, the fact is that I am not very religious. If I am correct in presuming that human space exploration is motivated more by spirituality than by science, I suppose I should give up the pretense that there is much value in doing "space biology." Certainly that is the position held these days by most biologists, who vehemently oppose the space station. I take a different stand. I think it is sheer hubris to believe that I or any mere mortal could stop the pilgrimage to other planets. There is nothing, however, in the religious quest that forbids atheists and agnostics from going along solely for the science. Let us try, then, to see that some real science, some high-quality science, gets done along the way.

*RICHARD WASSERSUG is a dual American and Canadian citizen and is based at Dalhousie University in Halifax, Nova Scotia.*