SEPTEMBER 1994
$3.95

SCIENTIFIC AMERICAN

Conquering Lyme disease.

The crisis in software.

What causes deep earthquakes?

***The past preserved**—a tomb painting copied by a member of Napoleon's army.*

SCIENTIFIC AMERICAN

September 1994 Volume 271 Number 3

Scientific American (ISSN 0036-8733), published monthly by Scientific American, Inc., 415 Madison Avenue, New York, N.Y. 10017-1111. Second-class postage paid at New York, N.Y., and at additional mailing offices. Canada Post International Publications Mail (Canadian Distribution) Sales Agreement No. 242764. 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; fax: (212) 355-0408 or send E-mail to SCAinquiry@aol.com.

DEPARTMENTS

THE COVER painting portrays a scene copied from the tomb of Egyptian pharaoh Ramses the Third, who reigned from circa 1198 to 1167 B.C. The precise rendering is one of many illustrations in *La Description de l'Égypte,* a text compiled by members of Napoleon Bonaparte's Commission of Science and Arts. These engineers and scientists accompanied the French army when it invaded and occupied Egypt between 1798 and 1801 (see "The Scientific Importance of Napoleon's Egyptian Campaign," by Charles C. Gillispie, page 78).

THE ILLUSTRATIONS

Cover painting reproduced courtesy of the Rare Book Division, Department of Rare Books and Special Collections, Princeton University Libraries

Page	Source
34	Russell C. Johnson, University of Minnesota
35	Roberto Osti
36	Roberto Osti (*left*), Russell C. Johnson (*right*)
37	Roberto Osti
38	John Radcliffe Science Photo Library, Photo Researchers, Inc. (*left*), Mark S. Klempner, Tufts University School of Medicine (*center*), Robert T. Schoen, Yale University (*right*)
39	Ruth R. Montgomery, Yale University
40–41	CERN
42–43	AIP, Niels Bohr Photo Library (*top left*), Argonne National Laboratory (*top center*), European Organization for Nuclear Research (*top right*), Ian Worpole (*bottom*)
44–45	Ian Worpole after Andrew Boden/Fermilab Experiment 771 Collaboration (*top left*), Ian Worpole (*all others*)
46	Cornell University; color manipulations by Laurie Grace
47	CERN
49	© 1994 C. Bruce Morser
50–51	Photograph courtesy of Alcoa (*top*), Steven Stankiewicz (*bottom*)
52	Johnny Johnson (*chart*), Steven Stankiewicz (*inset*)
53	Archive Photos
54–55	Dana Burns-Pizer
58	Jared Schneidman/JSD
59	Dana Burns-Pizer
60	Jared Schneidman/JSD (*top*), John Hartwig/ Harvard Medical School (*bottom*)
61	Jared Schneidman/JSD (*left*), courtesy of Thomas P. Stossel (*right*)
62–63	Jared Schneidman/JSD
65	Roberto Osti
66	U.S. Geological Survey
67	Laurie Grace
68	Harry W. Green II
69	Harry W. Green II (*bottom left* and *right*), Ian Worpole (*all others*)
70	Laurie Grace
71	Harry W. Green II
73	*Providence Journal-Bulletin/Mercury*
74–75	Johnny Johnson
76	P. Vauthey/Sygma
77	National Aeronautics and Space Administration
78–85	Rare Book Division, Department of Rare Books and Special Collections, courtesy of Princeton University Libraries
86–87	Courtesy of Denver International Airport (*top*), John Sunderland/ *The Denver Post* (*bottom*)
88	Laurie Grace
89	Katherine Lambert
90	Guy Marche/FPG International
91	Laurie Grace
92	Johnny Johnson
93	Katherine Lambert
94	Photograph courtesy of National Institute of Information Technology, New Delhi
95	Laurie Grace
104	Michael Goodman
106–107	Kathy Konkle

SCIENTIFIC AMERICAN®

Established 1845

EDITOR: Jonathan Piel

BOARD OF EDITORS: Michelle Press, *Managing Editor;* John Rennie, *Associate Editor;* Timothy M. Beardsley; W. Wayt Gibbs; Marguerite Holloway; John Horgan, *Senior Writer;* Kristin Leutwyler; Philip Morrison, *Book Editor;* Madhusree Mukerjee; Corey S. Powell; Ricki L. Rusting; Gary Stix; Paul Wallich; Philip M. Yam

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

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

PRODUCTION: Richard Sasso, *Vice President, Production;* William Sherman, *Production Manager; Managers:* Carol Albert, *Print Production;* Janet Cermak, *Makeup & Quality Control;* Tanya DeSilva, *Prepress;* Carol Hansen, *Composition;* Madelyn Keyes, *Systems;* Ad Traffic: Carl Cherebin; Kelly Ann Mercado

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

ADVERTISING: Kate Dobson, *Associate Publisher/Advertising Director.* OFFICES: NEW YORK: Meryle Lowenthal, *New York Advertising Manager;* Randy James, Rick Johnson, Elizabeth Ryan. CHICAGO: 333 N. Michigan Ave., Chicago, IL 60601; Patrick Bachler, *Advertising Manager.* DETROIT: 3000 Town Center, Suite 1435, Southfield, MI 48075; Edward A. Bartley, *Detroit Manager.* WEST COAST: 1554 S. Sepulveda Blvd., Suite 212, Los Angeles, CA 90025; Lisa K. Carden, *Advertising Manager;* Tonia Wendt. 235 Montgomery St., Suite 724, San Francisco, CA 94104; Debra Silver. CANADA: Fenn Company, Inc. DALLAS: Griffith Group

MARKETING SERVICES: Laura Salant, *Marketing Director;* Diane Schube, *Promotion Manager;* Wendy Robinson, *Research Manager;* Nancy Mongelli, *Assistant Marketing Manager;* Ethel D. Little, *Advertising Coordinator*

INTERNATIONAL: EUROPE: Roy Edwards, *International Advertising Manager,* London; Vivienne Davidson, Linda Kaufman, Intermedia Ltd., Paris; Karin Ohff, Groupe Expansion, Frankfurt; Barth David Schwartz, *Director, Special Projects,* Amsterdam. SEOUL: Biscom, Inc. TOKYO: Nikkei International Ltd.; TAIPEI: Jennifer Wu, JR International Ltd.

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

SCIENTIFIC AMERICAN, INC.
415 Madison Avenue, New York, NY 10017-1111
(212) 754-0550

CHAIRMAN AND CHIEF EXECUTIVE OFFICER: John J. Hanley

CO-CHAIRMAN: Dr. Pierre Gerckens

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

DIRECTOR, ELECTRONIC PUBLISHING: Martin Paul

CHAIRMAN EMERITUS: Gerard Piel

PRINTED IN U.S.A.

LETTERS TO THE EDITORS

Time Travel

In "The Quantum Physics of Time Travel" [SCIENTIFIC AMERICAN, March], David Deutsch and Michael Lockwood state that trips into the past do not violate any of the known laws of physics. They base this statement on the "many universes" interpretation of quantum mechanics.

Nevertheless, a review of their explanation and diagram reveals that in fact their time traveler violates a number of conservation laws. In disappearing from the B-universe and appearing in the A-universe, the time traveler certainly must carry the electrons in her body from B to A, thus violating the conservation of lepton number in both universes. In addition, she carries her mass and energy from B to A, violating the conservation laws of mass and energy. If she carries an electric charge, then electric charge is not conserved either.

Perhaps it could be argued that these conservation laws are obeyed only when all the alternative universes are taken into account. Unfortunately, this leads to conservation laws that may not be obeyed in any single universe and are therefore completely unlike those we now know.

Publish this letter. Otherwise I shall send it to you again last year!

ROBERT H. BEEMAN
Coral Springs, Fla.

What about Occam's razor? Complexity should not be added without good reason. Deutsch and Lockwood postulate the existence of uncountable parallel universes (a "multiverse"). That is one interpretation of the meaning of quantum mechanics, but it is not the only one, and we are not necessarily forced to accept it. Moreover, it does not explain anything real: no time-travel paradox has ever been known to occur, there are no actual indications of parallel universes and no time loops have ever been encountered.

A. R. PETERS
Enschede, the Netherlands

The authors attempt to eliminate the time-travel paradox by allowing travel only between parallel universes. In other words, time travel within a single universe is still prohibited. If one cannot travel into one's own past, how can it be said that one is traveling into the past at all?

LIONEL D. HEWETT
Chairman
Department of Physics
Texas A & M University

Deutsch and Lockwood reply:

Does time travel violate conservation laws? No. The laws of quantum physics, including conservation laws, do not in general determine events in a single universe but only in the multiverse as a whole. In our time-travel examples, no mass, charge or other property is ever created or destroyed. It merely travels from one place to another, perhaps in another universe.

Occam's razor properly applies to concepts, not universes. To say that there are "many universes" is no more than to say that big things obey the same physical laws that experimental physicists routinely apply to subatomic particles, which involve multiple trajectories or histories. What does violate Occam's razor is the introduction of additional elements—such as hidden variables or a collapse of the wave function—for which there is no experimental or theoretical justification beyond a stubborn attachment to a classical worldview.

Is what we described really travel into the past or just travel into another universe? Call it what you like, but if the terms "past" and "future" are to mean anything, they should refer to something physically observable. Therefore, if yesterday in "our" universe qualifies as the past, then so must yesterday in a universe that was physically identical to ours, even if it subsequently diverged.

Eco-Label Confusion

We appreciate being mentioned in "How Green is My Label?" ["The Analytical Economist," SCIENTIFIC AMERICAN, May], but the description of our Environmental Report Card by Marguerite Holloway and Paul Wallich is likely to leave your readers confused. The Environmental Report Card is not a seal of approval, nor is it viewed as one by consumers. In fact, it was developed precisely to overcome the observed deficiencies of the seal programs, through research and through input from government agencies, industry, and consumer and environmental organizations. It has earned praise from a wide range of environmental and scientific experts and is supported by major retailers.

Unlike seal programs, the Environmental Report Card does not set arbitrary standards to define what makes a product "green." Instead it presents the environmental burdens of a product in a straightforward manner. Every product, no matter how green, has some environmental burdens; the less energy and fewer resources used and the less pollution and solid waste created, the better. Companies are free to use any technology or process to reduce the burdens associated with their products, rather than being confined to a set of select technologies.

LINDA BROWN
Vice President, Communications
Scientific Certification Systems
Oakland, Calif.

Holloway and Wallich reply:

We did not say that the report card is a seal of approval, rather that consumers can *interpret* it as such. Nowhere does the report card state that it is *not* a seal of approval. Brown may not feel such a disclaimer is necessary. But when a consumer is faced with two products, only one of which bears a report card (in green ink), who could blame him or her for thinking that the graded product is somehow more benign? Furthermore, the label is hardly simple: the rating system is not based on readily accessible standards and does not ease comparisons between products with disparate environmental impacts.

Letters selected for publication may be edited for length and clarity. Unsolicited manuscripts and correspondence will not be returned or acknowledged unless accompanied by a stamped, self-addressed envelope.

ERRATUM

The caption on page 99 of "Nurturing Nature" [April] misidentifies the photograph at the left. It shows a mangrove wilderness.

SEPTEMBER 1944

"The war has led to the construction of many large flying fields well adapted to military needs, but has not produced a coordinated system of airports adequate for the real needs of the United States. There are now 3000 civil airports. Soon after the war there will be need for at least 3000 extra fields."

"If war-necessitated industrial plant construction has done nothing else, it has brought home forcefully the fact that clean plants, attractively designed, tastefully landscaped without and decorated within, are worth the slight extra cost and trouble that these features entail. Community pride is developed thereby and workers are happier."

"White-hot sheet steel moving 20 miles an hour as it emerges from a rolling mill can have its thickness accurately measured by x-rays. This new development is described as follows by Dr. William D. Coolidge, General Electric Vice-President in charge of research: 'X-rays may be used as a gauge without making mechanical contact with the work. With an x-ray outfit below and an x-ray intensity measuring device above the sheet, it becomes possible to have a constant indication of thickness and, if desired, to have the x-rays themselves control the mill so as to maintain automatically a constant thickness of the steel sheet.' "

"A series of studies have led A. R. Lauer, associate professor of psychology at Iowa State College, to conclude that unrestricted driver licenses should be given only to those having 'at least 20/40 vision in both eyes, or 20/30 vision in one eye. When vision reaches 20/80 or 20/100 it may be best to limit the applicant to daylight driving or to speeds below 30 miles an hour.' "

SCIENTIFIC AMERICAN

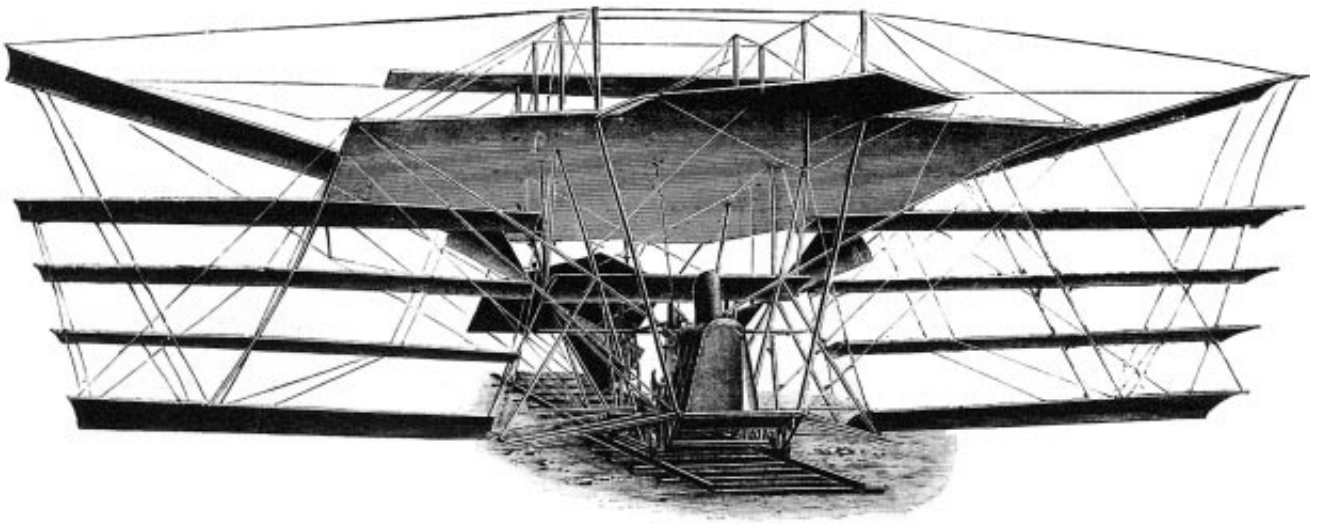
SEPTEMBER 1894

"The French War Office seems to be the target for all inventors, intelligent and otherwise. One invention takes the form of a captive shell, made to explode over fortresses, etc., and containing a small camera attached to a parachute. The enemy's fortifications would be photographed instantaneously, the apparatus hauled down like a kite, and the only remaining operation would be to develop the plates. Another inventor thinks that explosive bullets filled with pepper would have the twofold result of blinding the enemy and fostering French trade with its colonies."

"As the result of elaborate investigation, Dr. J. S. Haldane arrived at the conclusion that in colliery explosions the deaths from suffocation were due, not, as generally supposed, to carbonic acid gas, but to the preponderance of nitrogen and the deficiency of oxygen. Life could be saved if the colliers could be supplied with oxygen for an hour or so; and he has devised and exhibited an apparatus for enabling a man to breathe oxygen, of which 60 liters were compressed into a one-half liter bottle, with tube and regulating taps."

"In the department of dentistry the Chinese have anticipated by centuries the profession in Europe and America in the insertion of artificial teeth. A section sawed from the femur of an ox is utilized to fill the vacant space in the mouth. Through holes drilled in each end, copper wires are passed to fasten the bone to the adjoining teeth."

"On Tuesday, July 31, for the first time in the history of the world, a flying machine actually left the ground, fully equipped with engines, boiler, fuel, water and a crew of three persons. Its inventor, Mr. Hiram Maxim, had the proud consciousness of feeling that he had accomplished a feat which scores of able mechanics had stated to be impossible. Unfortunately, he had scarcely time to realize his triumph before fate interposed to dash his hopes. In a moment the machine lay stretched on the ground like a wounded bird with torn plumage and broken wings. Its very success was the cause of its failure, for not only did it rise, but it tore itself out of the guides placed to limit its flight, and for one short moment it was free. But the wreck of the timber rails became entangled with the sails, and brought it down."



The Maxim flying machine

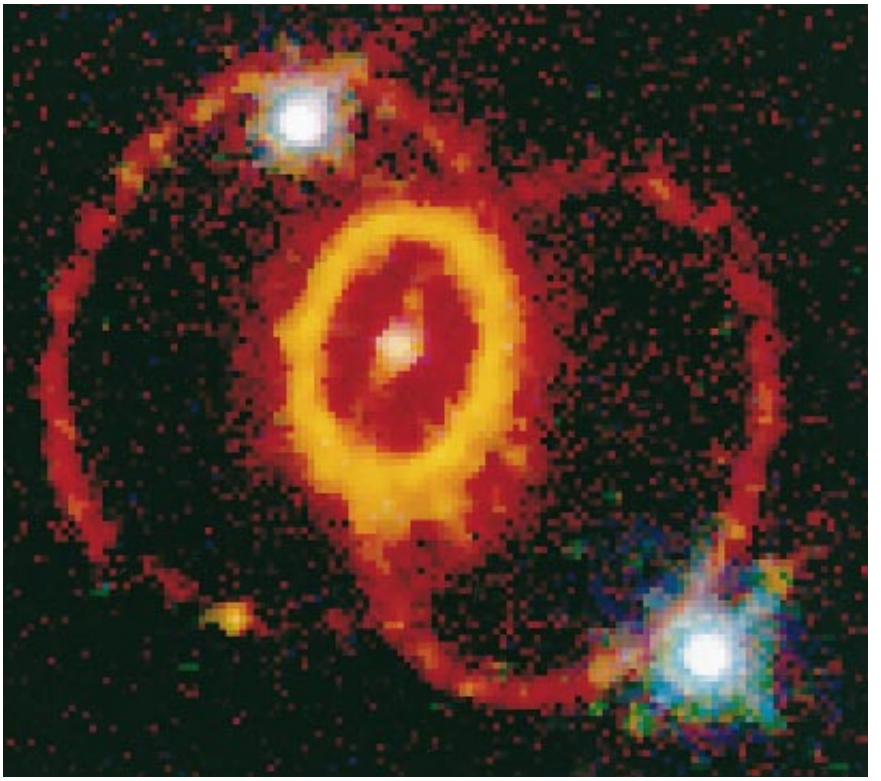
Super Loops

Strange, delicate rings of light frame a recent supernova

Nature has an astonishing ability to create grace out of devastation. The latest case in point is supernova 1987A, a blue giant star that dramatically obliterated itself seven years ago. A new view from the *Hubble Space Telescope* reveals three delicate, well-formed rings that have appeared around the exploded star. The image has both delighted and baffled astronomers. "It's beautiful—I even have it on a T-shirt!" exclaims Richard McCray of the Joint Institute for Laboratory Astrophysics in Boulder, Colo. But how could those rings have formed? "I'm stumped," he confesses. "There is nothing else like it in the sky."

Hints of the supernova's loopy nature began to emerge in 1989, when ground-based telescopes detected a bright ring. At first, researchers thought they had a good explanation for that celestial hula hoop, notes Christopher Burrows of the Space Telescope Science Institute, who conducted the latest *Hubble* observation. Some 30,000 years before its demise, the star expanded into a red giant star that puffed off a thick cloud of gas concentrated along its equator. Several thousand years ago that red giant evolved into a smaller, hotter blue star that emitted a wind of high-velocity gas. The blue-giant wind overtook the older, denser material and compressed it into a thin, hourglass-shaped shell. The brilliant flash of the supernova illuminated the dense waist of that shell, which appears as a ring.

McCray and his colleague Douglas N. C. Lin of the University of California at Santa Cruz now question that model, primarily because it is hard to understand why astronomers clearly see a narrow ring but find no hint of the other parts of the shell. Also, the ring is expanding far more slowly than one would expect from the above scenario. McCray and Lin propose instead that the ring is the inner edge of the flattened disk of gas from which the star formed several million years ago. If so, then astronomers are seeing, in a single snapshot, traces of the star's birth as well as its death.



NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

***RINGS OF GLOWING GAS** around supernova 1987A defy easy explanation. The large rings lie in front of and behind the bright inner ring, implying that these features are part of a tilted, hourglass-shaped structure.*

The origin of the faint outer loops around the supernova is even more obscure. Burrows offers a tentative explanation. He proposes that an unseen neutron star or black hole lies close to the supernova remnant. That star could shoot out twin, opposing jets of material that compressed two circular parts of the shell around the supernova; those circular parts, when struck by radiation from the exploded star, light up, producing the dual outer loops. McCray objects that Burrows's model violates "the tooth fairy rule"—a credible theory can invoke a mysterious, unknown agent ("tooth fairy") only once. But he agrees with Burrows that, for now, there is no better explanation.

Fortunately for scientists, supernova 1987A is not standing still. Debris from the explosion is racing outward; sometime around 1999 it will collide with the inner ring, giving rise to some spectacular millennial fireworks. The duration of those fireworks will reveal whether the ring is part of a thin shell or the inner rim of an extended disk, as McCray and Lin suggest. Furthermore, a spreading ellipse of illumination from the energized inner ring will gradually expose the outer rings and other currently invisible features in the region. The resulting three-dimensional picture of the supernova's surroundings will unfold "like a movie," McCray explains. Given the coming attractions, this looks like a show not to miss. —*Corey S. Powell*

Gone with a Bang

Supernova explosions create a gang of stellar runaways

Pulsars are among the strangest stars in the Milky Way. They are as massive as the sun but measure only about 10 kilometers across. They spin up to hundreds of times each second; during each turn, a pulsar's magnetic field whips up a pulse of ra-

diation that sweeps by the earth (hence its name). Now Andrew G. Lyne and D. R. Lorimer of the University of Manchester find that unlike normal stars, pulsars often do not even remain in the galaxy where they originated.

For more than two decades researchers have known that pulsars move faster than normal stars. New observations reveal the disparity to be much greater than workers realized, however. Last year a group headed by James M. Cordes of Cornell University observed the glowing trail of a runaway pulsar plowing through a gas cloud. Cordes's team estimates the pulsar travels at least 800 kilometers each second—so fast that it will break free of the Milky Way's gravitational clutches.

The study by Lyne and Lorimer demonstrates that such runaway pulsars are the rule, not the exception. The two workers examined a number of improved surveys of the apparent motion of pulsars across the sky. They also took into account recent work by Cordes and Joseph H. Taylor of Princeton University, which indicates that pulsars are systematically more distant than previously thought (which in turn implies that old estimates of pulsars' rate of motion were too low). In the end, Lyne and Lorimer concluded that the average pulsar is born traveling at a rate of about 450 kilometers a second, so fast that "about half of the neutron stars probably escape the Milky Way," Lyne says.

Earlier surveys had tended to overlook the fastest pulsars because their paths carry them out of the galaxy and away from the viewer, making them relatively faint and hard to detect. Those wayward stars form a giant halo around the bright spiral disk of the Milky Way. Many of the stars in that halo continue outward into intergalactic space, surrounding our galaxy with a vastly distended mist of neutron stars. Likewise, some of the old neutron stars now in the Milky Way may have originated in other galaxies, Lyne points out.

The discovery of runaway pulsars has inevitably raised the question of what accelerates these stars to such tremendous velocities. Most astronomers infer that a slight asymmetry in the initial supernova explosion sends the neutron star shooting away like a pinched watermelon seed. But at present, theorists cannot generate anything more than "hand-waving arguments" to explain how such asymmetries might come about, Lyne notes. (Theoretical modeling of supernovae has been sufficiently crude that, until recently, computer simulations routinely produced duds that collapsed instead of exploding.)

Uneven emission of neutrinos or ejection of gas during a supernova explosion could give pulsars the "kick" that explains their high velocities, reports Adam Burrows of the University of Arizona. Indeed, increasingly elaborate computer codes indicate that some such irregularities *must* occur during the explosion. Current models produce pulsar velocities that are considerably too low, however. "We haven't been able to put everything together yet," Burrows says. "The data show that there's a lot more violence than we've been able to simulate."

If the core of the exploding star receives a mighty shove in one direction, the supernova should also produce a lopsided cloud of debris. Robert A. Fesen and Kurt S. Gunderson of Dartmouth College may have detected such a feature in Cassiopeia A, the remnant of a supernova that occurred just 300 years ago. The two astronomers see a jet of gas racing away from the center of the explosion at 12,000 kilometers per second, twice the speed of the other parts of the remnant. "In at least one section, it was a very asymmetric explosion," Fesen concludes.

Even here, alas, the supernova story is far from clear. Observers cannot find

a pulsar connected with Cassiopeia A, and Fesen notes that there may be multiple jets pointing in various directions. Such features would further complicate the picture of what happens in supernova explosions. "This is not quite the smoking gun you're looking for," Fesen cautions. "The thing is smoking, but it's a bit cloudy." —*Corey S. Powell*

Sick, Sick, Sick

Neurotic? Probably, says DSM-IV

Do you use grammar and punctuation poorly? Is your spelling horrendous, and penmanship bad, too? You may be mentally ill—that is, if your diagnostician believes you are truly impaired and adheres strictly to the guidelines laid out in the latest edition of the *Diagnostic and Statistical Manual of Mental Disorders IV* (*DSM-IV*), published by the American Psychiatric Association. The manual lists these indications under Code 315.2, the "Disorder of Written Expression."

The *DSM,* or "the psychiatrist's bible," catalogues the behavioral traits associated with some 290 different psychoses and neuroses. The newest version is the third update published in the past 15 years, and critics charge that it shares a problem with its predecessors. "The criteria open a wide bag, and a lot of healthy people fall in," explains Herb Kutchins, a professor of social work at the California State University at Sacramento. Kutchins notes that tomboys could be diagnosed with gender-related personality disorders, or college students as alcoholics.

Kutchins and his colleague Stuart A. Kirk of the University of California at Los Angeles claim the book serves primarily as a guide to filling out insurance forms. "Most counselors use it for filing only, not for treatment planning or understanding clients better," Kirk says. To reach this conclusion, the two have polled social workers in the U.S. about how they use the *DSM.*

Allen Frances, chair of the psychiatry department at Duke University and chief author of *DSM-IV,* disagrees with Kutchins and Kirk. "They trivialize the very important role *DSM-IV* plays in clinical communication, treatment selection and facilitating research," he says. "Those of us who have worked on it for a very long time realize its limitations but also its enormous value."

Frances concedes that the guidelines do leave room for differences in clinical judgment. He points out, however, that no set of criteria could be strictly objective. "Criticism of the *DSM* system comes from people who consciously or unconsciously reify it," he says. "It's only when the criteria are taken too seriously or applied too literally that problems arise."

Such as finding that a large number of Americans are, well, a little off? Sadly, Frances thinks not. A recent survey done at the University of Michigan found that half of all Americans suffer during their lifetime from one or another of the illnesses in the *DSM;* a third are so afflicted in any given year. "The criteria are fairly common occurrences, and so a large number of the population will exhibit some of them," Kirk says. "What qualifies as a mental disorder is a complex question."

The 27-member revision committee behind *DSM-IV* tried to find an answer by conducting 150 research reviews, reanalyzing 45 data sets and performing 12 field trials. In the end, it weeded out all but eight new entries. Inhalant-induced anxiety disorder made the grade; minor depression did not. "There was not enough information to warrant its inclusion," Frances says. "We were concerned that simple and ordinary aches

and pains would be overdiagnosed."

Some preexisting categories were retested but none removed. In the past, gay activists lobbied to have homosexuality erased from the *DSM* register; feminists likewise had PMS banished to the appendix, awaiting further research. "Most diagnostic categories don't have opponents who demand that the APA scrutinize the evidence," Kirk says. "The arbitrary line of what gets included is drawn with some political sensitivity."

Still, the *DSM*'s contents must correspond to those found in the *International Classification of Diseases* (*ICD*), published by the World Health Organization. By treaty, the U.S. must base surveys of mental health on *ICD* standards. In some cases, more than one *DSM-IV* disorder falls under the same *ICD-IX* heading. And the *ICD-IX* numbers are different from those used in the *ICD-X,* which debuted last year. A *DSM* appendix explains how to cross-reference *ICD-IX* and *ICD-X* codes.

So why does *DSM-IV* use codes from an earlier version of *ICD?* "It may take another seven years before *ICD-X* standards are adopted in this country," Frances explains. By then, Kutchins ventures a guess that a new *DSM,* sure to be a publishing success, may be on the way. —*Kristin Leutwyler*

Hot Air

U.S. CO_2 emissions may put reduction goal beyond reach

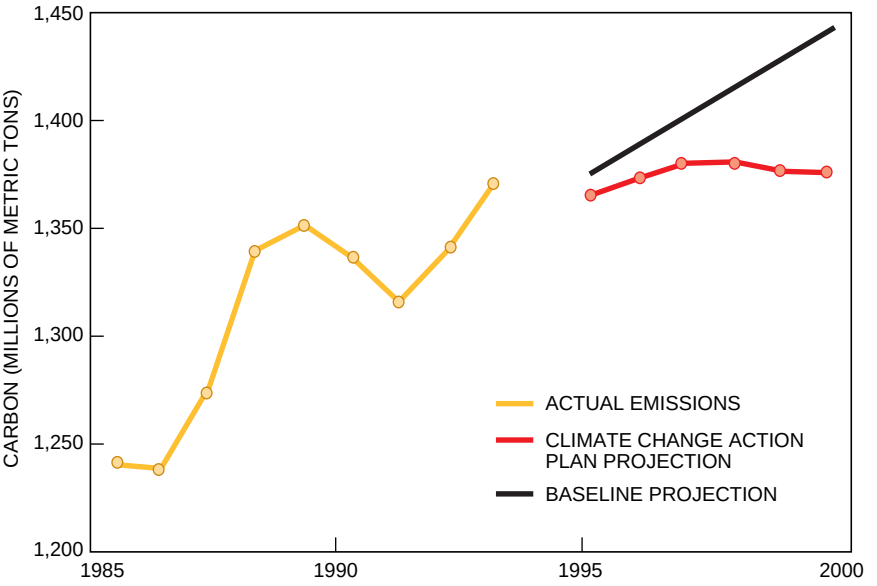
On April 21, 1993—Earth Day—President Bill Clinton announced that the U.S. would reduce its emissions of greenhouse gases to their 1990 levels by the year 2000. The pledge was intended to show that the U.S. took seriously the Framework Convention on Climate Change that had been agreed on at the Earth Summit in Rio de Janeiro in 1992. Other industrialized countries made the same promise. The administration followed through in October of last year by publishing its "climate change action plan," which specified how the target would be met.

Less than a year later the action plan is—if not quite in tatters—under severe strain. The document allows for an increase of 3 percent in U.S. carbon dioxide output by 2000 because emissions of other greenhouse gases are expected to fall, leaving a level total. But calculations completed in July by Howard Geller and Skip Laitner of the American Council for an Energy-Efficient Economy indicate that carbon emissions in the U.S. had by last year already climbed to 2.3 percent above the 1990 level, to 1,369 million metric tons.

The government's own carbon emission numbers will be published later this year, but officials say they are unlikely to differ significantly from Geller and Laitner's figures. Geller and Laitner used the Department of Energy's most recent estimates of 1993 fuel consumption. The calculation methods are standard. In other words, emissions have increased enough in three years to take up three quarters of the allotment for the whole decade. The U.S.'s commitment to return to the levels of 1990 by 2000 appears out of reach, unless strong new steps are taken to curb further growth in emissions.

Geller says the upturn in 1993 results largely from a 4.9 percent gain in economic activity since 1990. He and his colleagues as well as workers at the Natural Resources Defense Council have proposed several efficiency initiatives that they say could bring the target back in reach. The proposals include further improvements in automobile fuel efficiency and laws to require the use of recycled material in aluminum and plastic production. Geller's group would also like states to reform the regulation of utilities so that investments in energy efficiency will become at least as

U.S. EMISSIONS of carbon dioxide seem to be headed higher than those called for in the climate change action plan and higher than the baseline projection, which assumed no special controls.



profitable as those in energy supply.

If the U.S. fails to honor its commitment, which is not legally binding, it will be unable to say that it was not warned. The World Energy Council, an industry organization, stated in a report called *Energy for Tomorrow's World,* which was published last year, that there was "no realistic possibility" that under current policies developed countries could meet the goal of returning to 1990 emission levels by the year 2000 [see "Turning Green," page 96].

Looking at the world as a whole, estimated carbon dioxide emissions from fossil fuels have decreased slightly since 1991, according to estimates by the Worldwatch Institute in Washington, D.C. But analysts agree that the explanation for the fall lies in the recession and, especially, the economic chaos in Russia and eastern Europe. The 1991 oil fires in Kuwait may also have contributed. So the downswing is unlikely to be permanent.

It is tempting to see a link between the slight fall in carbon dioxide emissions resulting from human economic activity and a slowdown in the rate of accumulation of atmospheric carbon dioxide from all sources between 1991 and 1993. But the link is tenuous, say Charles D. Keeling and Timothy Whorf of the Scripps Institution of Oceanography in La Jolla, Calif., who monitor carbon dioxide levels at stations at the South Pole and on Mauna Loa in Hawaii. They think natural processes, including the eruption of Mount Pinatubo in the Philippines in 1991, are largely responsible for the slower buildup of the gas between 1991 and 1993.

An "El Niño," a periodic global climatic disturbance, persisted during those years, and that anomaly may temporarily increase the ocean's uptake of car-

bon dioxide. Pinatubo threw dust into the stratosphere that caused cooling below and, possibly, increased precipitation. Keeling and Whorf speculate that those effects spur plants to take more carbon dioxide out of the atmosphere.

In any event, the go-slow was only temporary: this year the rate of carbon dioxide buildup measured at Mauna Loa picked up again and is at the high end of predictions based on known human emissions. The great experiment—how life will change in a high–carbon dioxide atmosphere—seems to be getting under way. *—Tim Beardsley*

Backfire

Could Prozac and Elavil promote tumor growth?

Some oncologists have begun to contemplate the disturbing prospect that two of their favorite agents, Prozac and Elavil, might be medical boomerangs. Episodes of severe depression occur three times more frequently in cancer patients than in the general population, and women are victims of depression more often than men. Prozac and Elavil can alleviate the depression that often accompanies breast cancer and other malignancies. Now there is disturbing evidence that the popular antidepressants may accelerate tumor growth.

Concern emerged two years ago when a group of Canadian scientists reported that rodents that were given Prozac and Elavil experienced an increase in the rate of growth of breast cancers and increases in the weight of other tumors. Recent work with antihistamines deepens the concern. The research team, led by Lorne J. Brandes, an oncologist at the University of Manitoba, has revealed a possible mechanism by which antihistamines and antidepressants may encourage tumor growth.

Antidepressants and antihistamines are closely related in function. Both block chemical messengers that are released by white blood cells known as mast cells. Antihistamines counteract histamine, which triggers allergic responses. Antidepressants generally function by blocking the reuptake of serotonin, a neurotransmitter that is important in the regulation of emotions. Because the chemical structure of serotonin is similar to that of histamine, antidepressants can also interfere with histamine by binding to its receptor sites.

Brandes and his colleagues have discovered a new receptor site in the family of enzymes known as cytochrome-P450. Cytochrome-P450 is involved in regulating cell metabolism, detoxification of the intracellular environment and cell growth. Brandes believes that both the antidepressants and antihistamines bind to the cytochrome-P450 receptor sites. The result, he suspects,

GERARD KWIATKOWSKI

LORNE J. BRANDES studied the progression of cancer in rodents that received antidepressant or antihistamine drugs in doses equivalent to those for humans. He observed accelerated tumor growth.

is the tumor cell growth that his group observed.

Brandes also points out that the chemical structures of Prozac and Elavil are similar to those of tamoxifen and its derivative known as DPPE. Although these compounds are used to treat certain forms of cancer (tamoxifen has been standard breast cancer therapy since the 1970s), both have been connected to tumor growth. Tamoxifen can promote uterine cancer in some women, and DPPE has been observed to cause tumor flares in some patients.

What shifts the drugs from cancer therapy to cancer threat? "Promotion of cancer growth does not occur at all dosages," Brandes states. "In the case of DPPE, high dosages are used for tumor prevention. Low dosages, however, seem to accelerate tumor growth."

Brandes describes this unusual pattern of response as a "bell-shaped curve" in which promotion of tumor growth occurs most significantly in the low- to mid-dose ranges rather than at the highest or lowest amounts. This pattern of cancer promotion at moderate dosages is of particular concern to Brandes. "Toxicologists have assumed for years that high doses of a drug cause cancer, and if they don't see a problem at the highest dosage, they don't look at lower ones." Brandes studied the low- to mid-dose range of antidepressant drugs. For example, the rodents received the equivalent of a human dose of one to four Prozac pills a day.

Critics point out that the experiments involved mice that had been given a carcinogenic substance known as DMBA or had been injected with active tumor cells. Douglas L. Weed, chief of preventive oncology at the National Cancer Institute, feels the study might not be readily applicable to humans because, he notes, people do not have their tumors injected. Determining what accelerates tumor growth in humans is more difficult than it is in animals because control conditions are harder to monitor in humans.

But Brandes sees an apparent double standard. "Drugs are screened in animals for their safety for human use. When drugs decrease cancer in rats, people are excited. Now we're showing that, at certain doses, these drugs accelerate tumor growth in rodents, and people say we'd better wait and see. You can't have it both ways."

Until the debate is resolved, what should users of antidepressant drugs do? Jimmie Holland, chief of the psychiatry service at Memorial Sloan-Kettering Cancer Center in New York City, offers words of caution. "Depression is a problem that needs aggressive treatment. Many breast cancer patients who should be recognized as depressed remain untreated. This issue may compound the problem by making people afraid to take medication that they need."

Brandes agrees that for some people there is no choice except to take these types of medication. "There's no question that Prozac is an excellent antidepressant drug. But I am worried about the use of these substances in cancer patients." —*Sasha Nemecek*

Lonesome Cowpokes

U.S. particle physicists are seeking distant venues

Congress's cancellation last fall of the Superconducting Super Collider (SSC) was, as David B. Cline of the University of California at Los Angeles puts it, "a gut-wrenching experience." The nation's particle physicists had pinned all their hopes on the giant machine, which would have carried the search for fundamental particles far into an uncharted realm. The only competitor, the Large Hadron Collider (LHC) at CERN, the European laboratory for particle physics near Geneva, will be much inferior in its ability to unveil exotic new objects.

But it has an undeniable advantage: it will be built. So after months of agony and discouragement American particle physicists have come out of mourning to put forward a sober and conciliatory program for participating in their science. A major part of the recent plan, drafted by a panel headed by Sidney D. Drell of the Stanford Linear Accelerator Center (SLAC), would call for Americans to overcome their instinct to go it alone and to join the LHC.

But the LHC will probably not be completed before 2005. For the interim, the panel asks for a "bump" of $150 million over three years to complete more modest but interesting domestic projects that the SSC had shoved out of the limelight. Judging from the response, Congress is ready to apply some balm. "There seems to be a perception that high-energy physics has borne more than its share of deficit reduction," an observer notes. Representative Sherwood Boehlert of New York, an implacable foe of the SSC, is one of three congressmen who introduced a bill in late June to authorize much of what the Drell panel recommended. But he staunchly denies any change of heart. At a projected $11 billion, he says, "the SSC was way over budget, way behind

schedule and eating into the base program. I have received hundreds of letters from the science community applauding my success [in killing it]."

Still, the ghost of the SSC continues to lurk, even as the Department of Energy (which oversees the affairs of particle physics) prepares to negotiate with CERN about U.S. participation in the LHC. One issue is the reliability of U.S. support. "We've never had a project canceled," says Christopher Llewellyn Smith, director general of CERN. "Our member nations take their commitments seriously." Americans and Europeans alike question whether the U.S. can be a reliable partner.

CERN, in which 19 nations participate, is currently negotiating the funding of a long-term plan through 2005, with detailed provisions through 1998. Fermilab, on the other hand, will not know its budget for 1995 until the House and the Senate pass the Energy and Water Appropriations Bill and it is then signed by the president—a process that could carry over well into fiscal year 1995.

About 500 Americans already work on experiments at CERN and have free use of the facilities. "Like other laboratories, we have an open-door policy," Llewellyn Smith says. The expectation was that Europeans would likewise work at U.S. laboratories; with the SSC being canceled, there is little hope that the favor can be returned. "We feel like interlopers," Cline comments. "We don't pay our bills."

To participate in any major way, the U.S. will have to contribute an amount commensurate with the number of scientists involved. How much, Llewellyn Smith will not say. Nor will Boehlert reveal what the U.S. is willing to put up: "You don't show your hand before you start the poker game." But the $400 million mentioned by the Drell panel is "certainly doable," he adds. "Nothing compared to the SSC."

Most of the Americans who were developing detectors for the SSC at universities and laboratories are already working on the LHC detectors. Although CERN has been very welcoming of the Americans and their expertise, some small cultural adjustments are apparently in order. At a recent conference, one European urged his Yankee colleagues to "leave their cowboy boots behind." As few U.S. physicists at CERN wear this native gear, it would appear that the reference is to the tendency of the uninhibited inheritors of Lawrence, Richter, Feynman, Gell-Mann and Wilson to speak up so often that they dominate the discussions.

Yet unless their participation in the LHC is placed on a firm legal and financial basis, it is hard to see how the Americans can other than tiptoe. Currently they are funded by a mixture of money allotted for the SSC's funeral and some scraped together from their home institutions. According to Frederick J. Gilman of SLAC, a third of the 198 physicists who were at the SSC site have already left high-energy physics. The worst is not over. Most of the SSC funeral money will be spent by the end of 1994, and the Drell panel bump will not kick in (if it does) until 1996. "1995 will be tough," Drell admits.

In this climate, high-energy physicists are being forced to redefine their goals. "The SSC had created a mood of very high expectations," says James D. Bjorken of SLAC. Experimenters are reconciling themselves to filling in details of the Standard Model of particle physics. "We now have the most beautiful set of data," says Melissa Franklin of Harvard University, who belongs to a group at Fermilab that recently saw evidence of the top quark. But few physicists believe that the data hide any surprises. An upgrade planned for the end of the century should allow Fermilab to pin down the properties of the top quark and the mass of the W boson. Experiments on charge-parity violation at the future B meson factory at SLAC do, however, hold out some hope of the unexpected.

Franklin plans eventually to move on to the LHC and to one of several new cosmic-ray experiments. Many particle physicists are trying to continue exciting research by looking for high-energy particles in cosmic rays or for neutrino masses or proton decay in low-energy experiments. Others are gravitating back to realms that they had left behind, such as quantum chromodynamics, a turf since occupied by nuclear physicists.

If the experimenters are despondent, theorists are even more so. "Without experimental data, we cannot make progress," says Yoichiro Nambu of the University of Chicago. "We need a breakthrough." There are few fresh ideas in the field; both technicolor and supersymmetry, the two candidates for extending the Standard Model, have their problems. Without experimental guidance, there is no way to extricate or replace them. "Everything I can think of to calculate has already been beaten to death," sighs a young researcher.

What will the next century bring? A hitherto unimagined particle? Currently the Higgs ("a three-billion-Swiss-franc particle," Cline quips) is the only entity the LHC expects to discover. But if scientists knew exactly what they would find (as funding agencies require them to), there would be no point in finding it, Bjorken notes. The thrill is that one never knows what might be out there.

Research has already started on a next-generation linear collider that would smash together electrons and positrons rather than the protons of the LHC. The collisions between the light particles should be cleaner and easier to tease apart. Innovative mechanisms to accelerate these particles have been proposed; much research will be needed to make them viable.

Japanese physicists are eager to build the collider, but such a venture will almost certainly be international. "I have a fantasy," Bjorken chuckles, "that the next machine will be set in the Australian outback, funded by rich South Asian nations of the 21st century." Wherever it is, the future for the American physicists will surely be a long commute. —*Madhusree Mukerjee*

STANFORD LINEAR ACCELERATOR CENTER

STANFORD LINEAR ACCELERATOR, where the* B *meson factory is to be built, is one of the last sites where experimental particle physics can be done in the U.S.

Borrowed Savagery

Interloping viral genes may cause lethal strep infections

Until a few months ago, a painful throat and fever were the worst that most people expected of streptococcal infections. Today streptococci have become the "flesh-eating bacteria" immortalized in lurid headlines like "KILLER BUG ATE MY FACE," courtesy of the *Daily Star,* a British tabloid. All sensationalism aside, however, many medical researchers and microbiologists are pondering whether some group A streptococci, after 40 or 50 years of relative clemency, are becoming more virulent.

"If you look at strep infection a century ago, it was a lethal disease," reflects Vincent A. Fischetti, a strep researcher at the Rockefeller University. "Whether the ones we are seeing now are similar organisms, new organisms or ones that were sequestered somewhere and are now coming back isn't clear."

Group A streptococci are diverse: they constitute more than 80 strains, and

Can I Buy You a Drink?

When asked about the effects of alcohol on erotic sensibility, the porter in *Macbeth* replies, "It [drink] provokes and unprovokes. It provokes the desire, but it takes away the performance." The first point, at least, is a given in the popular mind-set (and the second, spoken only in hushed tones).

Now comes a Finnish-Japanese study sure to reinforce an amorous male's hope that liquor is the quicker pick-her-upper. A group of investigators from Alko, Ltd., the Finnish state alcohol monopoly, the Åbo Akademi University and the Shiga University of Medical Science did the work. Their research, published in *Nature,* indicates that the ordinarily low testosterone levels in women rise dramatically one to two hours after imbibing spiked lingonberry juice. That finding generated some tabloid excitement because increases in testosterone and other androgens are thought to increase sexual interest in both men and women.

The team found that testosterone concentration in the blood plasma of the female subjects vaulted most sharply among those who were ovulating. In these individuals testosterone increased by about one third. Women taking oral contraceptives demonstrated an even bigger jump. They experienced up to a fourfold rise (because they began the experiment with a lower baseline: the pill increases the level of estrogen and progesterone and thereby reduces the relative concentration of testosterone). In contrast, male subjects and women taking a placebo showed no elevation in their levels of testosterone.

"It was a surprise finding," says one of the investigators, C. J. Peter Eriksson, a biomedical researcher at Alko. "We were interested in the metabolism of alcohol and looking at hormonal effects, and this just came out." The workers do not know the precise cause of the rise but suspect the reason may lurk in the way women and men metabolize alcohol.

But with respect to why women (and men) report a stronger interest in an erotic encounter after a couple of shots, the study may constitute much ado about nothing. The researchers never investigated sexual response per se. "The thing to do is to measure behavior, to see if those changes coincide with hormone changes," points out Barbara B. Sherwin of McGill University, a psychologist who examines the interactions between hormones and behavior. She also questions the way testosterone levels were determined. "What is unfortunate is that they measured total testosterone," Sherwin says—unfortunate because little of the testosterone in the female body is active. Estrogen helps to create a protein that binds testosterone, so only a small percentage of the hormone actually circulates freely. "Unless it affects behavior, so what?" Sherwin remarks.

MERRY ALPERN

***COME HERE OFTEN?** A recent study finds that a couple of drinks raises the level of testosterone in women. The hormone is thought to be responsible for the libido.*

"It's unlikely that the magnitude of the testosterone change observed would have a major effect, given that sexual arousal is determined by so many different factors," says Jack G. Modell, a psychiatrist at the University of Alabama at Birmingham. Modell specializes in the behavioral effects of alcohol. He cites studies showing that people respond to alcohol according to their expectations about what the compound is supposed to do: if one believes it arouses, then it usually does.

Modell also observes that alcohol tends to be consumed in settings that lead to sexual encounters. Perhaps most obviously, alcohol can break down inhibitions. "Intoxication can be a convenient excuse to do what you want to do," Modell opines. At least until circumstances invite performance. *—Philip Yam*

each strain may have several clonal types. Of the 20 to 30 million cases of strep estimated to occur in the U.S. every year, fewer than 15,000 fall into the serious category of invasive infections. These can manifest themselves in a variety of life-threatening ways, including a devastating pneumonia and a syndrome resembling toxic shock.

About 10 percent of the invasive infections result in the "flesh-eating" condition called necrotizing fasciitis, which starts when aggressive strep bacteria colonize a break in the skin. The streptococci and the toxins they make can gradually spread throughout the body, destroying the surrounding flesh at the rate of an inch an hour. Approximately 30 percent of those people who develop the fasciitis—between 300 and 500 people in the U.S. annually—die, usually because they do not seek medical attention quickly enough.

Most physicians and researchers became aware of the fasciitis and the other invasive forms of the disease only within the past decade or so. "I was at a meeting about 10 years ago when a physician from South America told me he was seeing people with lethal strep infections who were dying within four or five days," Fischetti recalls. "He asked me whether I'd ever heard about such cases, and I said no." Clusters of similar infections were later reported in Sweden, Finland, Czechoslovakia, New Zealand, Canada, Great Britain and elsewhere, including the U.S.

Experts disagree about whether invasive infections are new and on the rise. Infectious diseases do routinely wax and wane, for reasons that are not always clear. For example, scarlet fever was formerly a fairly common and deadly outcome of strep infections. Antibiotics have certainly contributed to its near disappearance, but as epidemiologists have noticed, the incidence and the severity of scarlet fever were declining years before antibiotics were introduced.

Reliable epidemiologic data are sometimes hard to obtain because physicians in the U.S. are not required to report cases of strep to the Centers for Disease Control and Prevention (CDC). Nevertheless, statistics from local health authorities and some multistate studies do suggest that the incidence of strep has been creeping up. "That has led us to some very serious discussion here about whether we should begin a program of aggressive monitoring for strep in this country," remarks Bob Howard, a spokesman for the CDC.

If the invasive infections are a recent phenomenon, what might explain strep's sudden virulence? One guess is that the organisms have acquired new genetic information—and new characteristics—from viruses. As Fischetti says, "It's common for bacteria to pick up genes by being infected by a bacteriophage," a type of virus that can incorporate its DNA into that of a bacterial host.

P. Patrick Cleary, a microbiologist at the University of Minnesota, has found evidence supporting that hypothesis. He and his colleagues at the World Health Organization's strep reference laboratory have determined that some clones of the M1 strain of group A strep, which is associated with about 40 percent of the recent invasive infections, seem to have recently acquired genetic material from a phage. In that material is a gene that encodes a toxin called a superantigen, which according to Cleary is widely believed to be the cause of the strep-related toxic-shock syndrome.

Cleary also has another study, now in press with the *Proceedings of the National Academy of Sciences,* which shows that group A strep organisms can sometimes invade human epithelial cells. Even more exciting, he says, is that the virulent form of M1 strep is particularly adept at this intracellular trespassing. The clinical significance of this ability is still unknown, but it is a characteristic of many bacteria that can cause blood infections, such as salmonella and plague bacilli. Cleary doubts that the superantigen could be helping the streptococci enter cells, so the phage may carry a second gene that confers this ability.

Investigators are also still trying to determine whether the virulent strains of strep produce unusual quantities of enzymes such as proteases, which digest proteins, and hyaluronidases, which dissolve the substance that holds tissues together. Such molecules could be at work in necrotizing fasciitis. Phages sometimes carry genes for such enzymes, Cleary notes, adding that "this phage could be like a pistol loaded with many shots."

Whatever the cause and origin of the virulent strep infections, the prospects for treating and preventing them remain excellent. When used early in an infection, Howard says, penicillin is still "exquisitely effective" against strep; there is no evidence that strep organisms are building up any resistance to it. Fischetti is also developing an oral vaccine that might offer protection against all strains of group A strep; he hopes to enter clinical trials in a year or two. For the moment, if you feel a strep throat coming on, swallow hard and be grateful that's all it is. —*John Rennie*

A Marriage of Science and Art

In his 1959 book, *The Two Cultures and the Scientific Revolution,* C. P. Snow deplored the cleaving of the humanities and the sciences into separate, antagonistic ways of intellectual and moral life. The marriage between Jeremiah P. Ostriker and Alicia Suskin Ostriker that same year argues that the breach is more apparent than real. Jeremiah is chairman of the department of astronomy and astrophysics at Princeton University and an influential cosmologist. Alicia is professor of English literature at Rutgers University and a noted poet and essayist.

"Snow had it wrong," Jeremiah reflects. "I think the two cultures he described are much more like one another than the ones that he ranked in between." Alicia agrees, noting the similar ways that ideas are created and tested. "First you know something intuitively and then you try to prove it," she says. "If it turns out you can't prove it, then it's wrong. Writing a poem is much the same; you try to find the right words, and if you can't, you didn't really know the poem."

She also hails the practical advantages of their literal marriage of science and art. "People often ask, 'Isn't it a strange combination of professions?' My answer is always: one, it is not uncommon, and, two, it makes perfectly good sense to be married to someone creative who is not in your field and therefore with whom you are not directly competing. There are poets married to other poets. I don't know how they do it. If I were married to another poet, I'd be dead!" Alicia laughs.

The Ostrikers have always lived in close yet distinct worlds. They both grew up in Manhattan—he on the Upper West Side, home to much of the New York intelligentsia, she in housing projects at the island's north end. They met in high school and dated while they attended college in Boston (Jeremiah was at Harvard, Alicia at Brandeis); they married during their senior year. Jeremiah performed graduate work on the stability of rotating stars at the University of Chicago under the famed astrophysicist Subrahmanyan Chandrasekhar, who helped to bolster his perfectionist tendencies. Alicia, meanwhile, continued her work in literature at the University of Wisconsin, where she raced through her Ph.D. in three years, "a record, I think."

MERRY ALPERN

***JEREMIAH AND ALICIA OSTRIKER** mine similar veins of creativity in their seemingly disparate fields.*

In 1964, after completing graduate school, both Ostrikers applied for positions at a variety of universities, including Princeton. Alicia received a rude life lesson in the form of a letter telling her that "as a glance at our catalogue might have informed you, our faculty here at Princeton is entirely male, therefore my reply to your query must be in the negative." Fortunately, she received an offer from Rutgers, where she has remained as a professor of English. Princeton did offer a position to Jeremiah; he accepted, joining one of the nation's most prestigious astrophysics departments. He, too, has stayed put ever since.

During his time at Princeton, Jeremiah has steadily expanded the physical scope of his research, from stars to galaxies to the universe as a whole. In the early 1970s he began to consider the dynamics of a rotating galaxy. Drawing on his graduate work, he recognized that a flat, rotating spiral galaxy, like a rapidly rotating star, could not remain stable, "so I realized that galaxies can't be like that." Ostriker teamed up with his Princeton colleague P. J. E. Peebles to make computer simulations of galaxies. They found that the galaxies remained stable only if they were surrounded by a spherical halo of unseen material, commonly known as dark matter.

The resulting 1974 paper was a landmark in establishing the now conventional view that the visible universe represents only a small fraction of what is really out there. It also demonstrated Ostriker's ability to look past the common wisdom. "In young scientific fields, if you say all the accepted positions are wrong, you'll seldom be wrong," he offers as a kind of motto.

That attitude continues to guide his current work, much of which centers on developing and testing models that explain the origin of cosmic structure—galaxies, clusters of galaxies and the ever larger conglomerations that unfold in the latest maps of the universe. "There is no doubt that there was a big bang and that it was a hot big bang. But

do we understand the origin of structure within the big bang? I think the answer is no. My own guess is that none of the models we are looking at now is correct," Ostriker concludes.

He is particularly skeptical of the standard models of inflationary cosmology, an aesthetically appealing and popular elaboration of the big bang. Those models require that the universe contain a great deal more matter than observers can see or can deduce from the motions of galaxies. Furthermore, for theoretical reasons, that additional mass must consist of particles that do not interact with light or with ordinary matter. Unlike the dark halos surrounding galaxies, there is no direct evidence that the additional, exotic dark matter invoked in most inflationary models truly exists. "This is material for which there is no measurement but you wish there were," Ostriker says heatedly. "There is no reason other than ideology to have this."

So he and Peebles have been exploring cosmological models that do away with exotic dark matter. Ostriker has also continued to investigate the implications of cosmic strings, a cosmological constant, and other woolly astrophysical hypotheticals. "The way I think about them is that they are toys. My inclination is to play with them, insofar as there's some science in it, and I can use them to make sense out of some things that I couldn't otherwise. Who knows? One of them might be right," he says.

Alicia Ostriker sees eye-to-eye with her spouse on this issue. New ideas in literature are much like new ideas in astrophysics, she argues. "You test them against reality as you perceive it, and your work is a quest for truth." Drawing on that principle, she has evolved a literary voice as distinctive and freewheeling as her husband's style of research. In her first year of graduate school, a visiting professor lightly dismissed her poetry with the comment, "You women poets are very graphic, aren't you?" She credits that remark with goading her into thinking about what it means to be a woman poet. In her seven volumes of published poetry, Ostriker probes into many facets of that identity: sexuality, mortality and, above all, the physical experiences of the body.

After her second pregnancy, Ostriker began a lengthy poem on the experience of carrying and birthing a child. It appeared in 1970 under the title "Once More Out of Darkness." She was startled when a group of militant feminist students objected to her endorsement of motherhood. Undaunted, she wrote an essay in which she praised child rearing for putting the artist in touch with the factual world and issued a hope that mothering might one day enjoy a prominence in literature equal to that of sex and war.

As a critic, Ostriker began by writing a book on the radical British poet William Blake and editing a volume of his collected poems, an outgrowth of her graduate work. Then, during the 1970s, she turned to examining the nature of the female voice in modern literature in essays and collections, including *Writing Like a Woman* and *Stealing the Language.* In 1986 Ostriker's writing changed direction again because "my interests shifted, and I spent a lot of time reading the Bible." A new book, which she describes as "a feminist reading of the Bible, a real page-turner," will be published by Rutgers this fall.

Ostriker asserts that science is important in her work, even if her poems are never explicitly about a scientific topic. "My mind is shaped by what I know of science and my awareness of the scientific outlook on reality, and I often use scientific metaphors," she says. There is also a link between her writing style and her husband's research style, she thinks. "It's kind of similar: he's a cosmologist without a school, and I write as a feminist poet and critic without toeing any ideological line or dogma. It is probably the proximity to a scientific point of view, along with my own skepticism and sense of material reality, that keeps me from taking dogmatic positions."

"New ideas in literature are much like new ideas in astrophysics," argues Alicia Ostriker.

Jeremiah Ostriker likewise appreciates the importance of a literary approach in his intellectual life. He looks back to a class taught by the poet Archibald MacLeish. Each week MacLeish would pick out a poem and ask the students to analyze it in any way that they found interesting. The secret to producing a good paper was to pick out an interesting line of attack. "It struck me that this kind of teaching was much more helpful to me as a budding scientist than most of my science courses, which were basically like solving crossword puzzles for which you can look up the answer in the back of the book. I felt then, and I feel now, that when the answer isn't in the back of the book, the people who are good at science are the people who pick interesting problems and who figure out the right things to look at. It's a totally different set of skills, don't you think?"

His appreciation for nonideological thinking and his wife's experiences in academia have led Jeremiah Ostriker to reflect on the role of women in astrophysics. "I've been inclined for some time to write a book saying that a large fraction of the most important contributions to postwar astrophysics have been made by women precisely because they are outsiders," he relates. He points to several examples, including Beatrice Tinsley's recognition of galactic evolution, Vera C. Rubin's evidence for dark matter in galaxies and Neta A. Bahcall's discovery of the clumping of widely separated galaxy clusters. In each case, "people who understood things 'knew' that this was impossible," he says. "But the women didn't know."

The Ostrikers describe another, less welcome similarity between poetry and astrophysics: both fields sound intimidating to the layperson. Weary of the response to his actual vocation, Jeremiah used to tell people at parties, "I make bombs." Alicia describes the reactions when she meets strangers on an airplane and tells them her profession: "They go into instant paralysis and say something like, 'I don't even know how to spell,' or 'My wife likes to read.' "

Why do so many people find the scientific and literary worldview so alien and incomprehensible? Jeremiah traces the difficulty back to Euclid and the mathematical language of science. "That is what always seems to me the most remarkable thing—that the physical world obeys mathematics. The fact that mathematics works is extremely mysterious. There's something uncanny about it that is very disturbing to most people who are not used to it."

Alicia concurs that the problem dates back to the very origins of scientific thought but points a finger at Plato. "I think Plato was responsible not only for science but for the hatred of science, in that he invented dualism, the notion that ideally we should experience our selves, our souls, our essences, as separate from nature. Many scientists and humanists want to see human beings and the human mind as separate from the rest of the universe. But insofar as real poetry and real science get done, they get done by people who, consciously or otherwise, are operating as part of the universe rather than separate from it." —*Corey S. Powell*

SCIENTIFIC AMERICAN

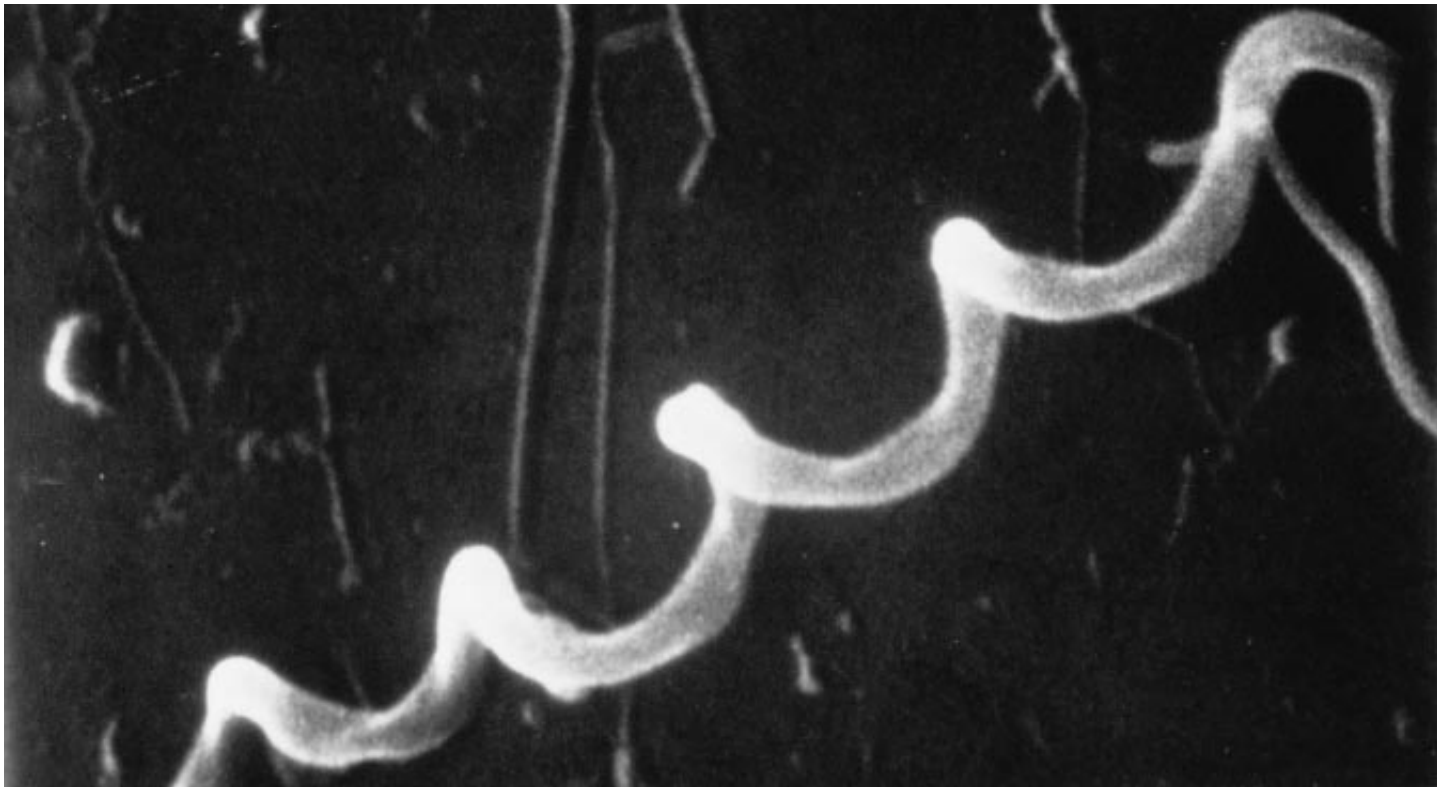
Disarming Lyme Disease

Antibiotics are usually curative. A vaccine is in clinical trials. Next on the research agenda: how to help people suffering from chronic symptoms

by Fred S. Kantor

Investigators first became aware of Lyme disease almost two decades ago. In rather short order, they identified the cause (a tick-borne microbe), showed that antibiotic therapy cures most cases and delineated the typical course of untreated disease. Recently the research has taken yet another heartening turn: a vaccine has been developed and is being tested in a large number of patients. At the same time, a difficult problem has moved to the forefront: Why does the disorder, which generally is self-limited, become chronic and occasionally debilitating in some patients? As the research enters a new stage, now seems an appropriate moment to summarize the insights gleaned during the first 20 years of study, to explain how the vaccine was developed and to highlight some of the latest thinking about the cause of chronic suffering.

Lyme disease was first recognized in Lyme, Conn. In 1975 two mothers were told their children had juvenile rheumatoid arthritis—a disabling condition in which joints become swollen and painful. The women soon learned that their children were not the only ones affected; many other children and adults in the region had been diagnosed with rheumatoid arthritis. This condition does not commonly occur in clusters, and so the mothers, in search of an explanation for the outbreak, contacted investigators at Yale University.



By the late 1970s Allen C. Steere and Stephen E. Malawista of Yale found that many of the patients they studied were afflicted with a mysterious disease that could produce a variety of symptoms, including but not limited to joint swelling. The cause was apparently a microorganism transmitted by at least one species of tick, *Ixodes scapularis* (then called *I. dammini*), prevalent in grassy areas and woods in and around Lyme. In 1982 Willy Burgdorfer of Rocky Mountain Laboratories in Hamilton, Mont., identified the microbe as a spiral-shaped bacterial species that now bears his name: *Borrelia burgdorferi.*

With the disease-causing agent in hand, researchers soon confirmed growing suspicions that certain skin conditions and neurological syndromes known in Europe were in fact manifestations of Lyme disease. Since that time, workers have identified the disease in many parts of the world, including Australia, Africa and Asia. In the U.S. it appears in almost every state but is especially prevalent in the Northeast, in Minnesota and in northern California (where the tick at fault is *I. pacificus*). Last year an estimated 8,000 cases were reported nationally.

The potentially disabling character of Lyme disease certainly justifies concern and vigilance. Yet it seems to me that media attention has been excessive and that the public is inordinately frightened. Most of the time, Lyme disease is easily treated and does not progress to the chronic stage. Indeed, it probably causes severe long-term effects in less than 10 percent of untreated patients. Recent studies have shown that many people who think they have chronic Lyme disease actually suffer from other maladies.

FRED S. KANTOR is Paul B. Beeson Professor of Medicine at Yale University. Before turning his attention to a vaccine for Lyme disease, he spent many years studying anergy, allergy, autoimmunity and the genetic basis of the immune response. Kantor is also a pilot; he has been flying small aircraft since 1957.

People who contract Lyme disease do so after an infected tick attaches itself to the skin. As the tick starts to take in a meal of blood, *B. burgdorferi* spirochetes in its midgut begin to multiply. They then cross into the tick's circulation, migrate to the salivary glands and pass with saliva into the host's skin. Luckily for potential victims, a tick has to be attached to a human host for 36 to 48 hours before an infectious dose of *B. burgdorferi* will be transmitted. This fact is comforting to those of us in areas where Lyme disease is endemic; we can establish a strong first-line defense just by checking ourselves assiduously for ticks every day.

Most people who do become infected will ultimately display one or more symptoms. Early on, perhaps 60 percent of patients will notice a roundish rash called erythema chronicum migrans (ECM). Three days to a month after spirochetes enter the skin, these individuals will see redness at or near the site of the tick bite. The reddened area, which neither itches nor hurts, expands over time and may grow to measure several inches across. It also typically clears in the center as it enlarges, so that it comes to resemble a target. Some other patients probably acquire this "target," or "bull's-eye," rash yet fail to see it, especially when they are bitten on the back or in the crease between the upper thigh and the buttocks. In the absence of antibiotic therapy, the lesion usually disappears within several weeks but sometimes fades within days.

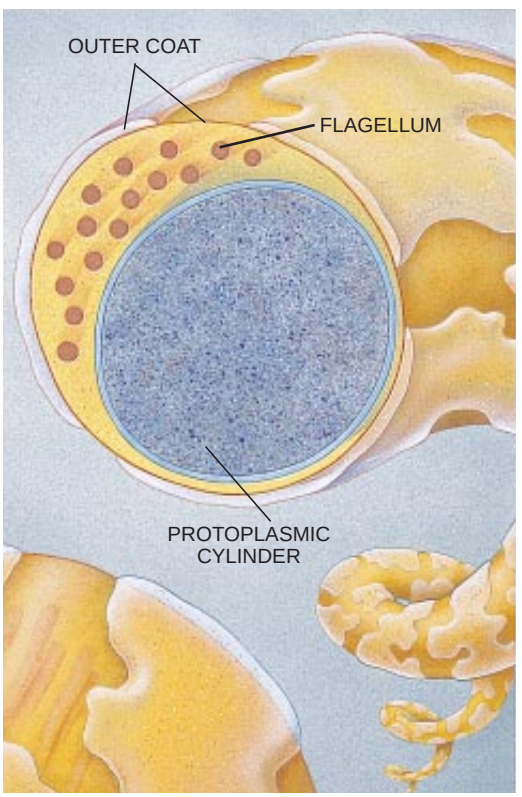
Days or weeks after a tick has introduced *B. burgdorferi* organisms into the skin, a variety of other fairly early symptoms, affecting many different areas of the body, may begin to emerge. These disorders are thought to stem from dissemination of the spirochetes to many tissues via the bloodstream. Flulike symptoms—chills, fever, fatigue, joint and muscle pains, and loss of appetite—arise frequently.

Early neurological problems appear often as well, in about 20 percent of untreated patients. In one such manifestation, called Bell's palsy, one or both sides of the face may become paralyzed for weeks or months before regaining full activity. Other early neurological symptoms can include meningitis (heralded by headache, stiff neck and sensitivity to light), encephalitis (which may cause sleepiness, memory loss and mood changes) and radiculoneuropathy. In this last condition, the roots of nerves that extend from the spinal cord to the periphery at some level of the body become irritated. Then the regions controlled by those nerves become painful and may tingle or go numb.

The heart is another organ that may be affected in the first weeks. The most common cardiac problem, evident in some 5 to 10 percent of infected individuals who go untreated, is atrioventricular block, a disruption in the heart rhythm. Most people will not be aware of this disturbance unless a physician detects it, although some patients will notice a decline in their ability to exercise. Fortunately, this condition tends to persist only for a week to 10 days and almost never requires insertion of a pacemaker.

Initial symptoms can also include mild musculoskeletal disturbances. Patients may have vague, migrating pain (but no swelling) in muscles, tendons or joints. Many people find that the temporomandibular joint is affected. Those symptoms generally diminish on their own over weeks or months. Nevertheless, about six months after the onset of infection approximately half of all individuals who have received no antibiotics suffer an episode of frank arthritis (marked by weeks of swelling and discomfort) in one or a few joints, particularly the knee.

In the U.S. an estimated 10 percent of untreated patients who suffer temporary arthritic symptoms of Lyme disease go on to acquire chronic Lyme arthritis.



MICROBE THAT CAUSES LYME DISEASE, the bacterial spirochete *Borrelia burgdorferi,* is shown intact (*left*) and in schematic cross section (*right*). Current research into vaccines is focused on inducing the human immune system to produce antibodies against a protein—outer surface protein A—in the outer coat.

***IXODES SCAPULARIS,* the tick that most often transmits *B. burgdorferi* spirochetes to humans in the U.S., is shown in its larval, nymphal and adult stages in the photograph (*left to right*). The adult depicted in the photograph is a female. Actual sizes of the unfed larva, nymph and male and female adult are indicated in the box (*top to bottom*). Ticks can become several times larger when they feed on the blood of a host.**

Patients with chronic arthritis may find that one or more joints repeatedly swell for months at a time or that certain joints remain enlarged and achy for a year or more. In contrast to many forms of arthritis (including rheumatoid arthritis), in which matching joints on each side of the body are affected, Lyme arthritis typically is not symmetrical.

In Europe, chronic arthritis is quite rare, but long-term neurological complications, such as cognitive deficits and dementia, have been documented in many patients. Moreover, up to 10 percent of untreated Europeans also suffer for years or decades with a disorder called acrodermatitis chronica atrophicans. In patients with this condition, affected areas of the skin become reddened and so thin and wrinkled that they resemble cigarette paper. In the U.S. these manifestations are rare. Variance in the frequency of certain symptoms presumably stems from differences in the strains of *B. burgdorferi* active in different areas of the world.

Sadly, cases of Lyme disease diagnosed late in the course of the illness may prove resistant to antibiotic therapy. Physicians sometimes prescribe repeated long courses of antibiotic therapy for patients with chronic disease. The value of this approach, which can have serious side effects (such as inducing the formation of gallstones), remains unproved, however.

A range of additional symptoms can appear at some point. Suffice it to say that dissemination of *B. burgdorferi* in the body can lead to disorders in practically every organ system, although the skin, heart, joints and nervous system are particularly targeted.

The exact molecular events leading to the symptoms of *B. burgdorferi* infection remain to be elucidated. Some evidence suggests they are caused by the body's own inflammatory response to microbes that have colonized target sites. During an inflammatory response, molecules and cells of the immune system (such as antibody molecules and macrophages) collect in infected tissue and attempt to eradicate any invaders. The inflammatory process can lead to swelling, redness and, at times, systemic effects, such as fever.

No matter what causes the manifestations of Lyme disease, the key to avoiding serious effects is prompt diagnosis and treatment of the underlying disorder. Regrettably, making a definitive diagnosis of Lyme disease during the early stages can be difficult, especially when the characteristic rash is not evident. The problem arises because various other symptoms, such as flulike complaints, can be caused by many other factors. Moreover, available blood tests for diagnosing Lyme disease detect antibodies that, in most cases, do not appear in the bloodstream until several weeks or months after the onset of infection. This property makes the tests unreliable for early diagnosis.

Investigators are working to develop alternative tests. Meanwhile many authorities recommend that no treatment be given for a tick bite alone. Physicians have to rely on their own judgment to determine whether Lyme disease is the probable cause of a patient's complaints.

Most people who become symptomatic do so in spring, summer or early fall. This pattern is now known to reflect peculiarities in the life cycle of *Ixodes* ticks. That cycle involves taking one blood meal (over the course of days) in each of three stages of development. *Ixodes* ticks have favorite hosts in every stage, but a range of animals, including humans, may be selected. In the case of *I. scapularis,* which accounts for most of the Lyme disease in the U.S., the larval form emerges in the summer from eggs deposited during the spring. It then attaches to a small vertebrate, typically the white-footed mouse (*Peromyscus leucopus*), and imbibes meal number one. If the host is infected with *B. burgdorferi* spirochetes, the larva that feeds on it can become infected as well.

Sometime after eating, the larva molts into a nymph. During the next spring and summer (mid-May through July), the nymph takes meal number two. If the larva was infected, the nymph will be infected and will thus be capable of transmitting infection to its host. The nymph, which before feeding is about the size of a poppy seed, accounts for most human infection. But it favors the white-footed mouse again, or other small vertebrates, as its food source.

By October the nymph molts into an adult the size of an apple seed. At that point, or sometimes in winter or spring, adults feed and mate to generate fertilized eggs and thereby initiate the cycle anew. Adults of *I. scapularis* often perform these activities on white-tailed deer (*Odocoileus virginianus*)—which explains why *I. scapularis* is often referred to as the deer tick. Deer do not themselves support colonies of *B. burgdorferi,* but they do carry ticks to areas where people live and play.

In the northeastern U.S., between 15 and 30 percent of all *I. scapularis* ticks, and some 50 percent of the adults, are infected. (Adult ticks are more likely to be infectious than nymphs because they have had an extra opportunity to feed on an infected host—once as a larva and again as a nymph.) In some places, such as Block Island and Nantucket Island, the figures are even higher. Even so, in most sections of the Northeast, only an estimated 1 to 3 percent of people bitten by *I. scapularis* contract Lyme disease.

The tick that transmits Lyme disease in California relies for its first or second meal on lizards or other hosts that are fairly resistant to infection by *B. burgdorferi.* Consequently, the rate at which ticks, and thus humans, are infected is much lower than it is in the northeastern U.S. The same is true of the species that transmit Lyme disease in certain regions of Europe and Asia.

About five years ago several of us at Yale began to wonder if we could devise a vaccine that would protect people from acquiring Lyme disease. Aside from me, the primary members of our group were John F. Anderson, Stephen W. Barthold, Erol Fikrig, Richard A. Flavell and Stephen Malawista. At the outset, we needed to satisfy ourselves that the human body could be induced to guard against colonization by *B. burgdorferi.* We took some encouragement

from the experience of people living in or near Montauk, N.Y., on the outer tip of Long Island. Before Lyme disease was recognized as an entity, people in the area frequently turned up with a condition dubbed Montauk knee. In what we now know is a manifestation of Lyme disease, one knee would swell and remain enlarged for many weeks before returning to normal. Anecdotal reports indicated that, once the condition went away, it did not recur. The lack of recurrence implied that the first infection induced the immune system to ward off new attacks.

Some animal work also suggested immunization was possible. Notably, Russell C. Johnson of the University of Minnesota injected hamsters with inactive *B. burgdorferi* organisms in the hope of inducing the immune system to react strongly against the foreign spirochetes. He then demonstrated that the hamsters could indeed fight off infection by living *B. burgdorferi* spirochetes that were injected into the animals later.

We began our experimental work by trying to determine which components on the spirochete best elicit a protective immune response. We paid special attention to proteins on the outer surface of the spirochetes, partly because surface molecules tend to be most accessible to the host's immune system.

By 1989 Alan G. Barbour of the University of Texas Health Science Center at San Antonio had cloned the genes for two such proteins—outer surface protein A (Osp A) and outer surface protein B (Osp B). From the Osp A gene, we synthesized a supply of Osp A protein and injected the molecules into mice. To our delight, the animals were fully protected against subsequent challenge by a large dose of *B. burgdorferi* spirochetes. We showed that Osp B could protect animals as well, although only if we exposed them to relatively low numbers of spirochetes.

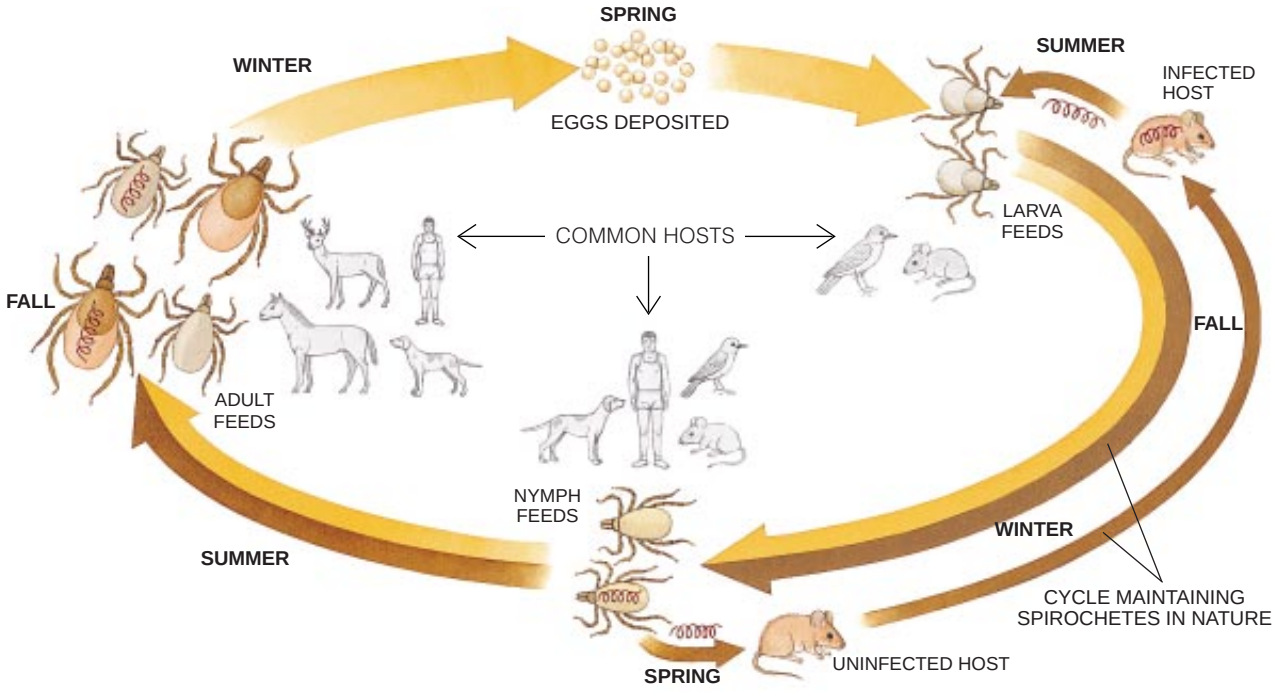
Later, with Jonathan Sears, then a medical student at Yale, we located the antigenic segment of Osp A, the part that evokes the immune response. It resides in the half of the molecule that is connected to the carboxyl (COOH) terminal of the protein. The research also revealed that the immune response induced by Osp A and Osp B was directed mainly by antibodies able to recognize and bind to these antigens.

Although the Osp A vaccine proved successful in our early studies, we had more work to do before we could consider embarking on human trials. High on the list was addressing a concern, expressed mostly by entomologists, that the immune response elicited by the vaccine might not provide protection against spirochetes that were injected by ticks instead of by syringes. Follow-up tests with ticks put this fear to rest—and gave us an unexpected result. Spirochetes had disappeared from the midgut of ticks that fed on vaccinated

The Life Cycle of the Deer Tick

The two-year life cycle of *Ixodes scapularis* (*thick arrows*) includes three feeding sessions. The cycle begins when females deposit fertilized eggs in the soil (*top*). By summer larvae emerge and imbibe a blood meal from a small vertebrate, usually the white-footed mouse. By the following spring or summer, larvae have molted into nymphs, which feed once more, typically on the mouse again. By fall, adults emerge and then, or somewhat later, feed a third time and mate, often on the white-tailed deer. Males die after mating; females, after depositing the eggs.

The tick's early preference for the white-footed mouse helps to maintain a related cycle (*dark arrows at right*)—one ensuring that *Borrelia burgdorferi* spirochetes (*spiral*) persist in the tick population. In this second cycle, larvae take up spirochetes from infected mice (*top right*) and molt into infected nymphs. Nymphs then pass the infection to more mice, which transmit it to larvae again, and so on. Lyme disease is most often transmitted to humans by nymphs that step out of this cycle and bite people. It can also be transmitted by adult ticks emerging from nymphs that became infected during their larval or nymphal stages (*dark arrow at left*).



rodents. Evidently, when the ticks took in blood from the treated animals, they also ingested anti–Osp A antibodies and other immunological substances that led to destruction of the spirochetes.

We immediately perceived that if we could immunize mice in nature, we would not have to inoculate people. Clearing of spirochetes from mice would quickly reduce the reservoir of infected animals. Then the number of infected ticks would drop, and the threat of Lyme disease might evaporate. Excited by this prospect and encouraged by some preliminary work, we tried to vaccinate mice by lacing their food with Osp A. To our disappointment, the animals did not become immune. Nevertheless, immunization of mice in the field remains a worthy goal.

We had to confront other concerns as well. The ideal human vaccine would protect people against all strains of *B. burgdorferi.* Would our vaccine, based on the Osp A protein derived from a single strain, be effective against other strains as well? In a large series of studies, we vaccinated animals and then challenged them with *B. burgdorferi* organisms isolated from ticks obtained from many parts of the country. No strains proved able to infect the vaccinated animals. This pattern held even on Nantucket Island, where *B. burgdorferi* isolated from wild ticks showed some variability in the antigenic region of Osp A. These findings gave us confidence that a single Osp A vaccine could shield people against infection by most strains of *B. burgdorferi* they would encounter in the U.S. The same vaccine might not be as effective in Europe, however, because the strains there are more diverse.

Since these studies were completed, *B. burgdorferi* variants that make highly mutated versions of Osp A have been found in the U.S. Some of these organisms produce abnormally short versions of Osp A or make a hybrid protein in which the antigenic region is replaced by a region normally found in an Osp B protein. Nevertheless, such mutants do not seem to be at all common in the field.

Osp A is not the only protein being investigated for its value as a vaccine, but so far it is the most promising. In the mid-1980s Thomas G. Schwan of Rocky Mountain Laboratories suggested that another protein on the outer surface of *B. burgdorferi* might be a vaccine candidate. This molecule, referred to simply as the 39-kilodalton protein (because of its molecular weight), was appealing because a fairly invariant version was found in strains of *B. burgdorferi* in different areas of the world. If the antibodies it elicited proved protective in people, a single vaccine would probably be usable worldwide. Yet we and others found no evidence that the molecule induces protective immunity.

Another surface protein purported to vary little from strain to strain—Osp C—also seems disappointing. An early suggestion that it could yield protection has not been confirmed. More recently, investigators have cloned the genes for three other outer surface proteins: Osp D, E and F. Unfortunately, none of the proteins evokes a strong protective response in vaccinated animals.

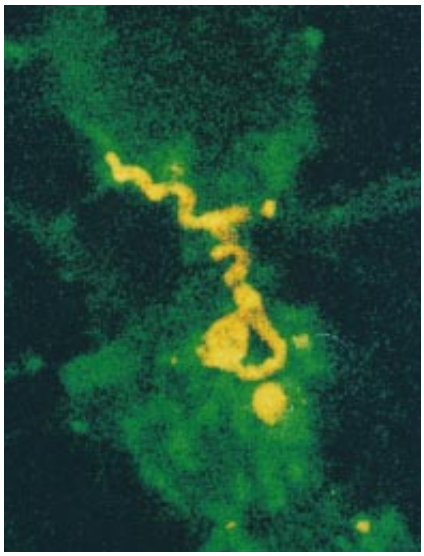
As efforts to find useful *B. burgdorferi* antigens continue, so do clinical trials of the Osp A vaccine. Two virtually identical versions are being evaluated, each produced by a different manufacturer. If all goes perfectly, at least one of them could be available in the U.S. by 1996. Patients living in Europe are participating in these studies; their experiences should give us an idea of whether a more diverse vaccine is needed on that continent.

At one time it was hoped that a vaccine capable of inducing a strong antibody response could also serve as an early treatment. Animal work has scotched that possibility, however. Animals injected with protective antibodies as soon as two or three days after exposure to infected ticks proved unable to resist the proliferation of *B. burgdorferi* spirochetes. Exactly why the antibodies failed as an after-bite treatment remains to be seen.

Even with ready availability of a vaccine and antibiotic therapy, a small fraction of people will undoubtedly continue to acquire infections that progress to the chronic stage. To help these individuals, researchers must first understand the events that give rise to the chronic state. One school of thought proposes that advanced disease stems from an autoimmune process. This view holds that *B. burgdorferi* spirochetes somehow induce the immune system to perceive one of the host's own proteins as foreign. Then the defensive system starts to attack normal tissue and keeps on attacking long after the spirochetes have been eradicated. The evidence in favor of the autoimmune explanation is not strong. Nevertheless, the fact that powerful antibiotics (which presumably kill spirochetes) can fail to eliminate symptoms gives some weight to the idea.

A contrary theory, with more data in its favor, postulates that chronic symptoms arise from the long-term persis-

SYMPTOMS OF LYME DISEASE can include an expanding, "bulls-eye" rash that often clears in the center (*left*)—the most characteristic, and usually the earliest, manifestation. Other common symptoms are Bell's palsy, in which one side of the face or both sides may be temporarily paralyzed (*center*), and swelling of one or a few joints, especially a knee (*right*).



***B. BURGDORFERI* spirochetes (*yellow*) are visible in this optical section through macrophages (*green*) taken from an infected mouse. The macrophages ingest and destroy *B. burgdorferi* organisms during the normal immune response. Some workers suspect the bacteria may occasionally settle into compartments where they are protected from immune attack. Such protected organisms could well contribute to chronic Lyme disease.**

tence of spirochetes. In other words, a subset of spirochetes continues to thrive somewhere in the body after evading normally effective immune defenses and, possibly, antibiotics. The coexistence of infection and a potentially curative immune response is termed concomitant immunity. The phenomenon is believed to underlie chronic disease in many illnesses, including some forms of cancer and parasitic disease.

If concomitant immunity were operating, one would expect to find an abundance of antibodies in the blood (a sign of an ongoing immune response) and evidence of spirochetes in the body (a sign of ongoing infection). As predicted, we and others have recovered antibodies against *B. burgdorferi* from the circulation of patients with recurrent symptoms. In addition, viable spirochetes, which are difficult to obtain from human tissue, have been isolated from the skin, joints and cerebral spinal fluid of some chronically ill patients who have high levels of circulating antibodies.

Indirect evidence that spirochetes are active in chronic disease comes from application of the polymerase chain reaction (PCR). This test amplifies small bits of DNA. It has revealed that *B. burgdorferi* DNA is present in the inflamed joints of some patients with chronic Lyme disease who have high levels of circulating antibodies against *B. burgdorferi.* To my mind, this finding suggests whole *B. burgdorferi* organisms are also present, because spirochete DNA would be unlikely to persist very long after the spirochetes that carried it had perished. It is nonetheless conceivable that the DNA is merely a footprint left behind by bacteria that have long since disappeared. Studies designed to distinguish between these two interpretations should resolve this issue soon.

How could *B. burgdorferi* spirochetes evade destruction by antibodies? One solution would be for the microbes to alter their own surface in ways that would make them invisible to the antibodies. They could, for instance, radically alter the structure of one or more surface antigens. There is ample precedent for such behavior in the microbial world, where organisms often revise the makeup of their coat after finding themselves in an inimical environment of antibodies. Yet, as noted earlier, *B. burgdorferi* does not seem much inclined toward making such changes. Moreover, spirochetes recovered from animals that have been infected for months or years are no different from the microbes that originally produced the infection—even in animals that have mounted a vigorous immune defense.

Still, there are other ways of altering the coat. Perhaps the organisms shed surface antigens, enticing host antibodies to interact with the free antigens instead of the pathogens themselves. Or perhaps the bacteria cover themselves with a host fluid or molecule, which the immune system then ignores.

Instead of disguising their outer surface, *B. burgdorferi* spirochetes could hide in places where they would be inaccessible to antibodies. An obvious refuge would be the inside of cells, where the cell membrane intervenes between the pathogen and antibodies. Indeed, one laboratory has found that *B. burgdorferi* can survive in macrophages—paradoxically, the very cell type that normally participates with antibodies in attacking *B. burgdorferi* organisms. During such an attack, macrophages ingest and degrade antibody-bound microbes. In the case of concomitant immunity, some spirochetes might find their way to a privileged compartment, shielded from the molecules that the cells deploy against ingested "prey."

Singly, neither major strategy—revision of the coat or hiding out—forms a completely satisfying explanation of how *B. burgdorferi* escapes destruction. Together, though, a combination of these and other tactics might well enable the bacterium to perpetuate itself, evading both antibodies and antibiotics. Ongoing studies will eventually reveal the precise maneuvers employed by the spirochete and suggest interventions.

Interestingly, all this research into Lyme disease may help improve understanding of syphilis, which displays many similarities to Lyme disease. The microbe that causes this sexually transmitted disease is another spirochete—*Treponema pallidum.* It, too, is capable of disseminating to many different kinds of tissues and causing chronic, antibiotic-resistant disease in some people. Further, many of the symptoms of syphilis resemble those of Lyme disease. Like *B. burgdorferi, T. pallidum* can cause skin rashes, cardiac abnormalities, nerve pain and dementia. Unlike *B. burgdorferi,* however, *T. pallidum* is very difficult to grow in the laboratory. As the molecular bases of infection by, and immunity to, *B. burgdorferi* emerge, researchers should gain new ideas for preventing syphilis and ameliorating its chronic effects.

In 1994, then, the emphasis of research into Lyme disease differs greatly from what it was in the 1970s and 1980s. The cause of the disorder and feasibility of prevention are no longer pressing questions. The big challenges are finding the optimal vaccine for each part of the globe, understanding the processes that perpetuate chronic Lyme disease and improving treatment for the late symptoms. With diligence and luck, perhaps these challenges, too, will be overcome quickly.

FURTHER READING

LYME ARTHRITIS: AN EPIDEMIC OF OLIGOARTICULAR ARTHRITIS IN CHILDREN AND ADULTS IN THREE CONNECTICUT COMMUNITIES. A. C. Steere, S. E. Malawista, D. R. Snydman, R. E. Shope, W. A. Andiman, M. R. Ross and F. M. Steele in *Arthritis and Rheumatism,* Vol. 20, No. 1, pages 7–17; January–February 1977.

PROTECTION OF MICE AGAINST THE LYME DISEASE AGENT BY IMMUNIZING WITH RECOMBINANT OSPA. E. Fikrig, S. W. Barthold, F. S. Kantor and R. A. Flavell in *Science,* Vol. 250, pages 553–556; October 26, 1990.

THE BIOLOGICAL AND SOCIAL PHENOMENON OF LYME DISEASE. Alan G. Barbour and Durland Fish in *Science,* Vol. 260, pages 1610–1616; June 11, 1993.

ANTIGENIC STABILITY OF *BORRELIA BURGDORFERI* DURING CHRONIC INFECTIONS OF IMMUNOCOMPETENT MICE. S. W. Barthold in *Infection and Immunity,* Vol. 61, No. 12, pages 4955–4961; December 1993.

Low-Energy Ways to Observe High-Energy Phenomena

By observing interactions that are forbidden in the Standard Model, physicists can peek at supersymmetric and other happenings

by David B. Cline

In the fall of 1993 Congress canceled the Superconducting Super Collider, or SSC. The SSC was designed to search for particles beyond the energy range of current accelerators. The Large Hadron Collider at CERN, the European laboratory for particle physics near Geneva, will probably be built in the first few years of the 21st century. But its energy is only about half of that which the SSC might have achieved. So how can physicists seek the massive particles that give logic and symmetry to theories of the fundamental elements of matter?

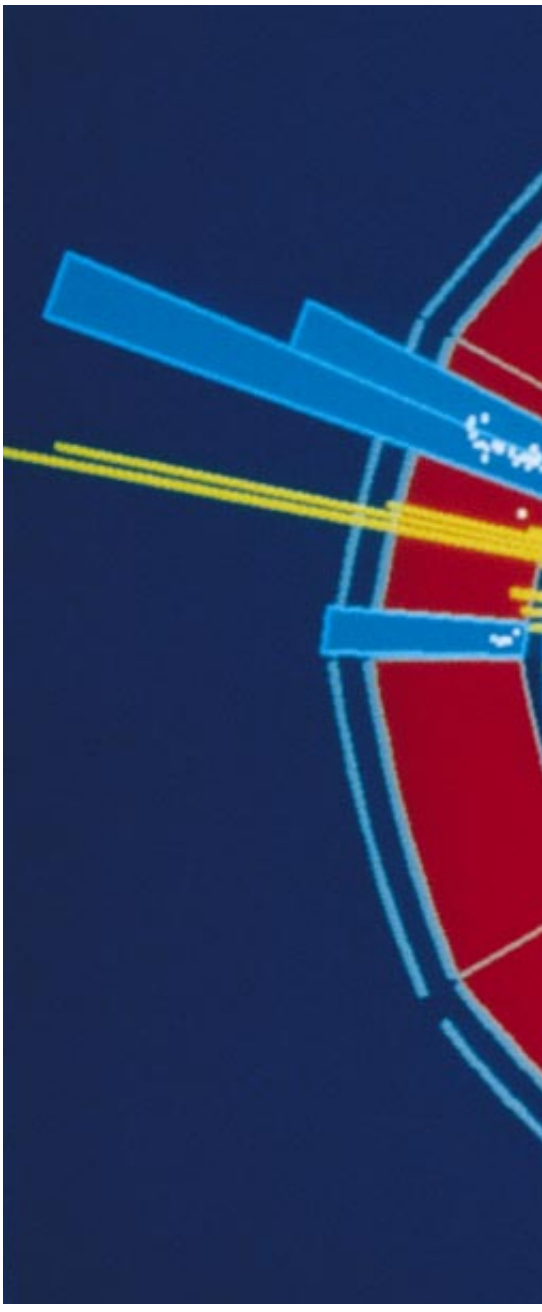
Fortunately, nature has provided a loophole through which scientists can look more deeply into its puzzles. Within the Standard Model of particle physics, some types of interactions are conceivable but in practice never seen. For example, a strange quark is not observed to decay into a down. Different means by which the interaction might occur manage to cancel one another out. Interactions that are not found to occur are said to be forbidden.

But it is entirely possible that particles not yet known to us might be able to mediate such an interaction by passing from one (known) particle to another. If researchers test ever more precisely, they may ultimately succeed in finding a faint signal for the process. Indeed, the detection will be made possible by the fact that the result one expects from the Standard Model is zero. Although it is difficult to discern a minute deviation from a large (and usually ill-defined) quantity, it is relatively easy to measure a deviation from zero. Once scientists have observed this so-called forbidden interaction, they will have evidence of the presence of a new particle. They can then add the particle to the Standard Model, thereby extending it.

One class of such interactions goes by the name of flavor-changing neutral currents, or FCNCs. Although these interactions had never been observed (until recently), new and exotic particles would almost inevitably create FCNCs that could be detectable in extremely sensitive experiments. Already this window may have revealed the first signs of particles that lie beyond the Standard Model.

Traditionally physicists have sought additional characters of the Standard Model by smashing together beams of known particles in accelerators. The mass-energy contained in these particles is oftentimes channeled into creating unknown ones. But the heaviest particles, which require large inputs of energy, are inaccessible to accelerators. In this realm, too, FCNCs have an advantage. As a rule, the heavier an exotic particle, the more likely it is to interact with a known one. Thus, although heavy particles are hard to generate in accelerators, they are easier to detect through their effects at low energies.

Known particles belong to the low-energy world that human beings normally live in. One class of particles com-



DAVID B. CLINE, a professor of physics and astronomy at the University of California, Los Angeles, helped to initiate the study of weak neutral currents in the 1960s. He participated as well in the discovery of the *W* and *Z* bosons in 1983. Weak interactions continue to be a primary interest. Another current activity is searching for proton decay at the Gran Sasso Laboratory in Italy. Cline is also designing an instrument at U.C.L.A. to test the sacred CPT theorem of particle physics, which states that the product of charge, parity and time reversal is conserved in particle interactions.

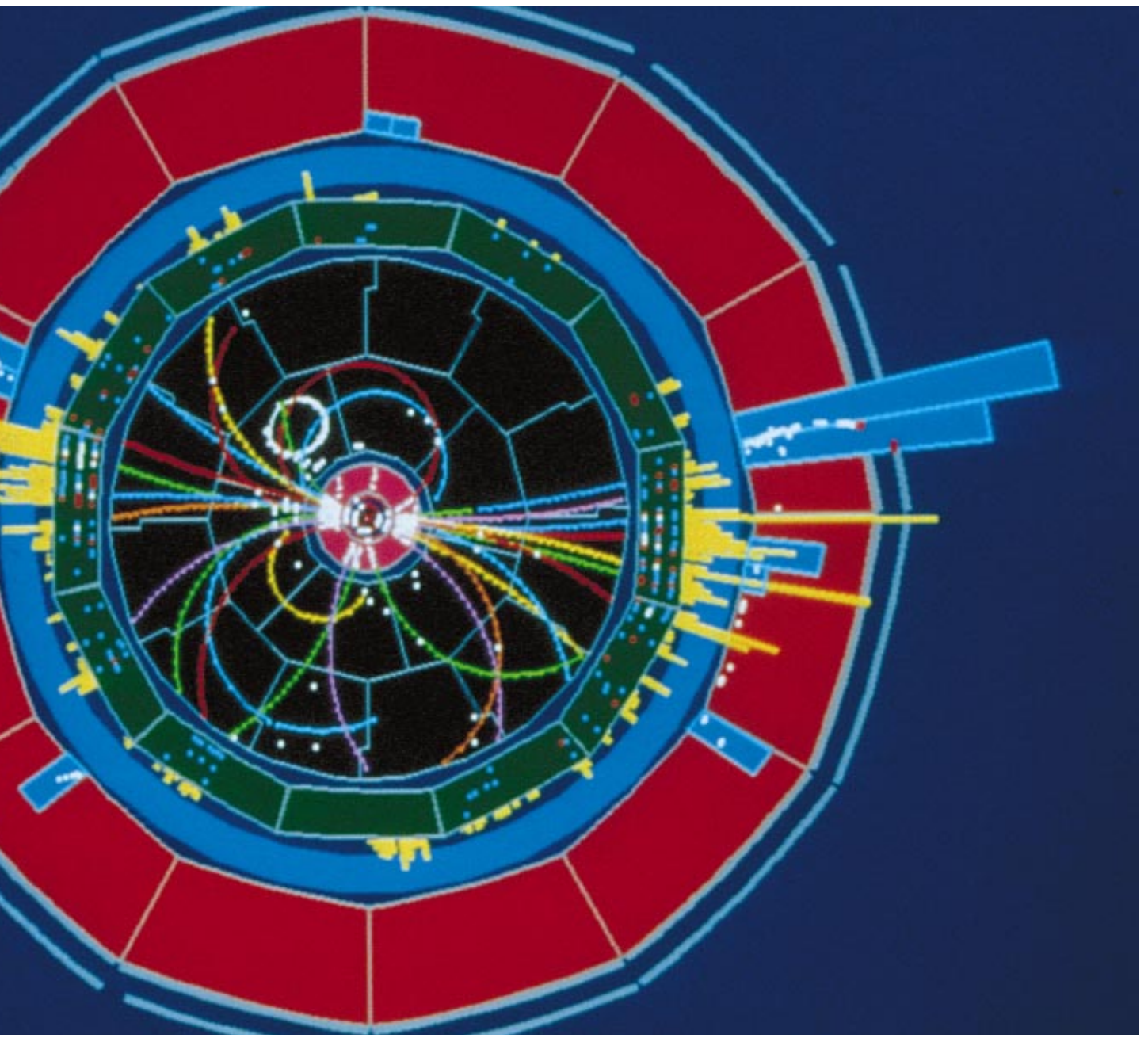
prises the leptons—electrons, muons and taus—and the elusive ultralight particles they decay into, the three neutrinos. Then there are the quarks.

Quarks seem to come in six types, or "flavors"—up, down, strange, charm, bottom and, now, top. Each quark is heavier than the preceding one in the list; the conservation of mass-energy allows a heavier quark to decay into one that is lighter, but not vice versa.

Up and down, strange and charm, and bottom and top are closely related to each other and are paired into "families." Up and down, for instance, are the two lightest quarks and belong to the first family. In each family one quark has an electric charge of $^2/_3$ (up, charm and top), and the other has an electric charge of $-^1/_3$ (down, strange and bottom). (The charge is measured in units of a proton's charge.) For every quark or lepton there is an antiquark or antilepton, which is identical except for having the opposite charge.

Quarks are able to change into one another by giving off or absorbing heavy particles. Three particles that transmit the weak nuclear force between quarks are the Z^0, the W^+ and the W^-. (The superscripts indicate electric charges of 0, +1 and –1, respectively.) For instance, a down quark can change into an up quark by a weak process, with the W^- particle carrying away the extra charge. Because the decay involves the passage of a charged particle (the W^-), it is said to be mediated by a charged current. Alternatively, a quark can interact with itself by emitting and reabsorbing a Z^0, which gives rise to a weak neutral current, or WNC.

DECAY OF A Z^0 PARTICLE is captured by the Aleph detector at CERN. The Z^0, which was first seen in 1983, transmits the weak force between other particles such as quarks, giving rise to a weak neutral current. Here it breaks up into a quark and an antiquark, which further splay into more stable particles such as mesons.



The History of Weak Interactions

The first inkling of a fourth force came in 1896, when Henri Becquerel observed that an atomic nucleus could decay by emitting an electron. By the 1930s it became evident that this "beta decay" involved the transformation of a neutron within the nucleus to a proton. In the 1970s physicists realized that a down quark in the neutron was changing into an up quark, forming a proton and emitting a W^- particle. The latter decayed into an electron and an antineutrino. The W^- and its relatives, the W^+ and the Z^0, mediate the weak force.

Enrico Fermi (*left*) first wrote down, in 1933, an interaction that described beta decay. Wolfgang Pauli (*right*) postulated that a new particle, a neutrino, carried away the extra energy in beta decay. Here they relax with their friend Werner Heisenberg at Lake Como in 1927.

1933

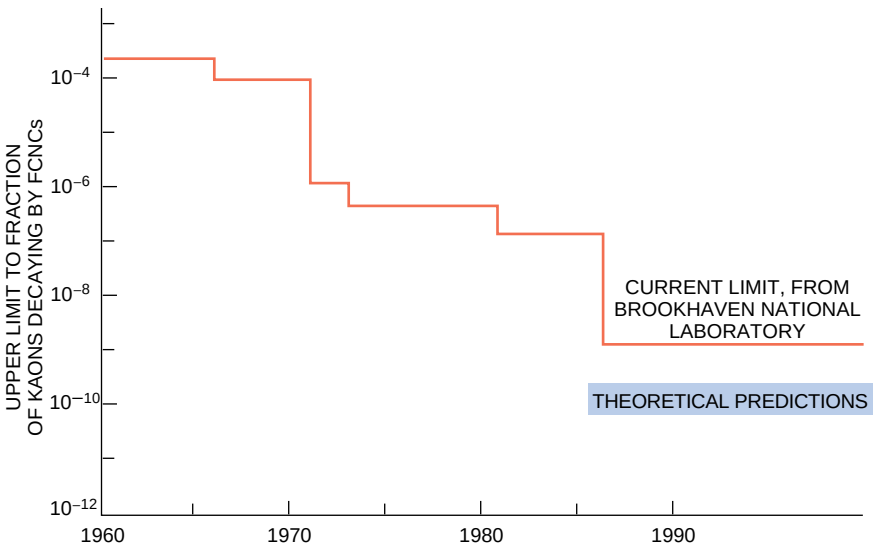
But never do experimenters see, as mentioned, a strange quark changing into a down, a process involving a flavor change. Because both these quarks have the same charge, such an interaction would have to proceed by a flavor-changing neutral current, or FCNC.

The absence of FCNCs in (almost) all experiments conducted to date has already led to the prediction—and discovery—of the charm and the top quarks. When physicists first became aware, in the late 1960s, that FCNCs did not seem to occur, they were at a loss to understand their absence. The theory of electroweak interactions had just been invented by Steven Weinberg, now at the University of Texas at Austin, and Abdus Salam of the International Centre for Theoretical Physics in Trieste, Italy. Previously Sheldon L. Glashow of Harvard University had described the same theory. They had fit the weak and electromagnetic interactions into the same framework and predicted the existence of the Z^0, W^+ and W^- particles. These particles became analogues of the photon, which transmits electromagnetic forces.

But the electroweak theory, brilliantly confirmed over the next decades, required the existence of neutral currents, in which a Z^0 is exchanged. Among other interactions, researchers assumed that the Z^0 might mediate the decay of the strange quark to the down. An experiment mounted at Lawrence Berkeley Laboratory in 1963, which I helped to initiate, did not find any such decays. What we did not realize at the time was that we were looking for a special, forbidden process: an FCNC. We simply concluded, on the basis of our experiments, that no neutral currents existed.

The only quarks known then were the up, down and the strange. In 1970 Glashow, John Iliopoulos of the École Normale Supérieure in Paris and Luciano Maiani of the University of Rome noticed that if a fourth quark existed, it could cancel the interaction of the strange quark with the down. Thus, the absence of FCNCs would be accounted for. Also, weak neutral currents that do not change flavor would exist. Because it would solve a long-standing dilemma, the theorists called their hypothetical fourth quark the "charm."

Meanwhile scientists at CERN and at Fermi National Accelerator Laboratory (Fermilab) in Batavia, Ill., had been looking for WNCs in processes involving neutrinos. Neutrinos interact with other particles only by weak interactions and with other neutrinos only by WNCs. For some time, different and confusing signals for WNCs from one of the major experiments led the physics community to claim, tongue in cheek, that "alternating neutral currents" had been discovered.



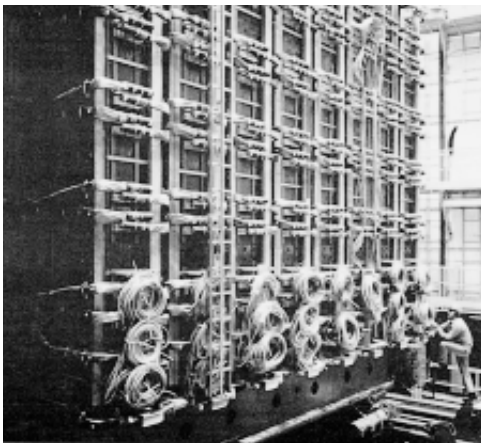
UPPER LIMIT to the fraction of kaons decaying into a pion (by emitting a neutrino and an antineutrino) has gone down steadily over 30 years. Fewer than one kaon in a billion decays in this way. The absence of this flavor-changing decay, involving the transformation of a strange quark into a down quark, led to the discovery of the charm quark and has restricted several extensions of the Standard Model. The most recent search is being conducted at Brookhaven National Laboratory.

In 1973 both the experiments at CERN and Fermilab found WNCs [see "The Detection of Neutral Weak Currents," by David B. Cline, Alfred K. Mann and Carlo Rubbia; SCIENTIFIC AMERICAN, December 1974]. In 1974, also at Fermilab, a charm quark made a fleeting appearance. Furthermore, large numbers of charm particles were produced in 1976 at the Stanford Linear Accelerator Center, thus confirming the theorists' scenario. Their formula for getting rid of FCNCs, called the GIM mechanism, has since turned out to have much broader validity than earlier envisaged. Within each family, one quark

Neutral current interaction occurs in a bubble chamber at the Argonne National Laboratory. An unseen neutrino, moving upward, initiates a chain of reactions culminating in a spiraling electron. This fundamental process was first observed at CERN and Fermilab.

1973

UA1 detector was built by an international collaboration for observing the carriers of the weak force. In 1983 it detected a *W* particle, earning a Nobel Prize for Carlo Rubbia, who was recently the chief of CERN.



1983

prevents the other from decaying via an FCNC.

Like the charm, the top quark was predicted to exist—because the bottom was not seen to decay to a strange or a down. Because each quark has a familial pair, FCNCs cannot easily occur within the Standard Model. Only on rare occasions can the heavy quarks violate the GIM mechanism, which works best for the light quarks.

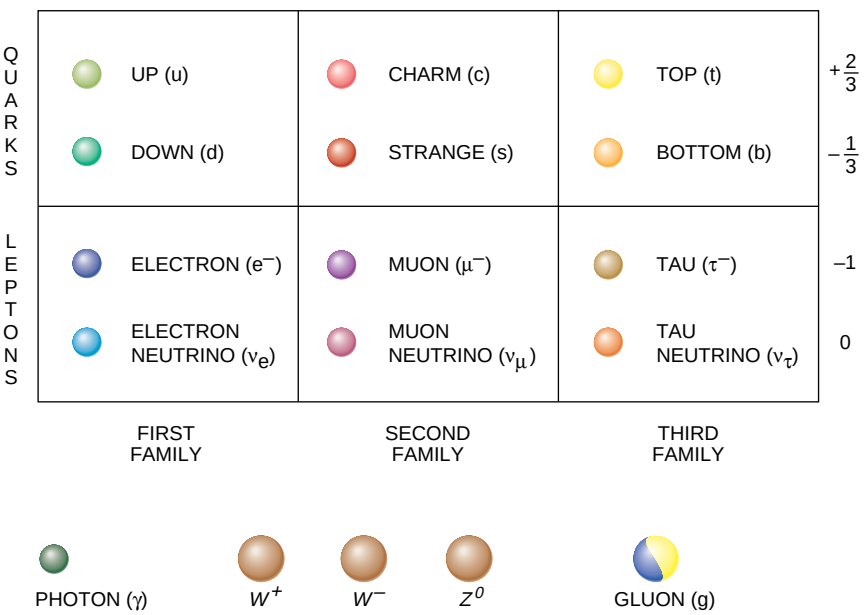
The rare FCNC that might be mediated by known particles—and, in fact, all particle interactions—is best illustrated by a kind of diagram invented by the late Richard P. Feynman of the California Institute of Technology [*see box on pages 44 and 45*]. In a Feynman diagram the particles are drawn as leaving traces, rather like a jet plane leaving a vapor trail. Thus, when two particles interact, their traces join at a vertex; when a particle decays, its trace breaks up.

An FCNC can occur if a top quark mediates the interaction in a way described by a complicated Feynman diagram known as a penguin. (The name has an unusual source. John Ellis of CERN once lost a game of darts with Melissa Franklin, now at Harvard. The penalty was that he had to put the word "penguin" into his next published paper—in which this diagram first appeared.) This decay, however, takes place infrequently, if at all. The penguin diagram has many variations; in most of them, exotic particles serve to mediate the decay.

Such particles are invariably postulated in theories that address the deficiencies of the Standard Model. One such problem is the question of why the fundamental particles have such diverse masses. The top quark, for example, is some 30,000 times heavier than the more common up quark, one of the principal constituents of ordinary matter.

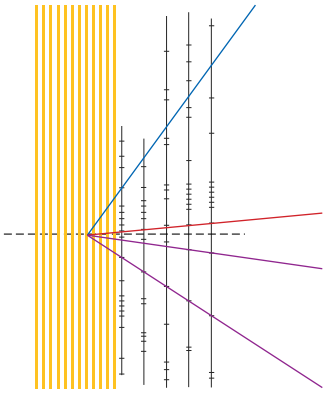
Particles are believed to gain mass by interacting with the heavy Higgs particle, which is also predicted by the electroweak theory. Because each quark has a different mass, however, it must couple with the Higgs with a different strength. These coupling strengths, or, alternatively, the quark masses themselves, are among the 21 parameters of the Standard Model that do not emerge from its fundamental assumptions. The properties have instead to be determined by experiment. This large set of arbitrary numbers is less than appealing—at least to those scientists who believe that at the deepest level of structure, the universe must be simple.

Theorists' prescriptions for tying up such untidy edges usually entail the prediction of yet more exotic and massive particles. One kind of extension of the Standard Model, for instance, is "grand unification." We have good reason to believe that at a very high energy the strong force (which holds the nucleus together) becomes unified with the electroweak. These forces become equally strong, joining to form a grand unified force. In that case, leptons become relatives of the quarks, and sev-

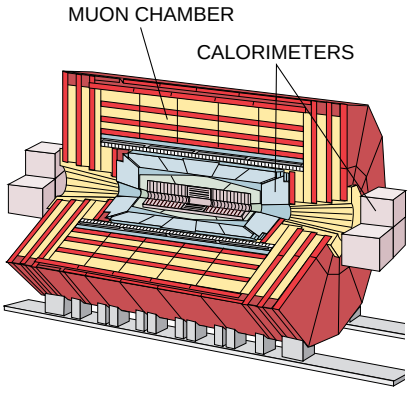


CHARACTERS OF THE STANDARD MODEL are the quarks and leptons, the photon (which mediates the electromagnetic force), the W^+, W^- and Z^0 particles (transmitting the weak force) and gluons (mediating the strong force). Each quark has a different flavor, but quarks and leptons in the same column belong to the same family. The numbers to the right indicate the electric charge of all particles in the same row. For every quark and lepton there is an antiquark or antilepton with the opposite charge. Quarks have another quantum number, called color, that has not been indicated. There are a total of eight gluons, each with a different combination of color quantum numbers.

A ***B*** **meson decays at Fermilab into a pion (*blue*), a kaon (*red*), and a muon and antimuon (*purple*). The meson, created by a proton coming from the left and striking a silicon wafer (*yellow*), cannot be seen. Copious sources of *B* mesons are the most promising locations for flavor-changing neutral currents.**



Compact Muon Solenoid may detect muons that signal a *B* meson's decay via a flavor-changing neutral current. This detector is to be used with the Large Hadron Collider at CERN, which is planned for 2003.



1991 2003

eral parameters relating to the strong forces become the same as those of the weak. The overall structure of a grand unified model is much simpler, and more rational, than that of the Standard Model. But it also requires the existence of ultraheavy particles, called grand unified particles, that have a mass of about 10^{16} GeV (1 GeV, roughly the mass of a proton, is a billion electron volts).

Among other interactions, these ultraheavy particles allow quarks to change into leptons—and the proton to decay. Physicists have looked for proton decays for more than a decade, and the searches are now becoming more definitive. With Carlo Rubbia of CERN and others in Italy, I am working on the ICARUS proton decay experiment at the Gran Sasso Laboratory in Italy. Giant detectors are being constructed at Gran Sasso and in Japan.

But there is a problem with the grand unified model. Its ultraheavy particles, by interacting with particles of the known world, would increase the masses of the latter. Quarks and leptons would then also have masses of about 10^{16} GeV. In that case, not only would humans not have observed them, but also they would not exist—at least in their current form.

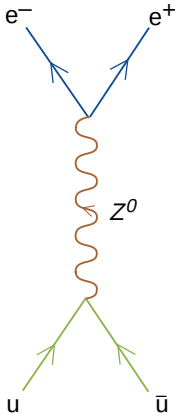
The only solution known to this "hierarchy problem" is supersymmetry, or SUSY. Supersymmetry postulates that each known particle is one of a supersymmetric pair. The superpartner of a quark, for example, would have a heavier mass and a different spin, or angular momentum. It would in effect cancel the interaction between the heavy grand unified particles and the quarks and leptons of the world, solving the hierarchy problem.

Many theorists are convinced that supersymmetric partners must exist. But none have been found. Maurice Goldhaber of Brookhaven National Laboratory sometimes jokes that the situation is not that bad: we at least have one half of all supersymmetric particles in the universe—the quarks and leptons!

One necessary consequence of super-

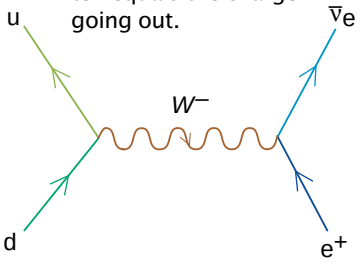
Feynman Diagrams

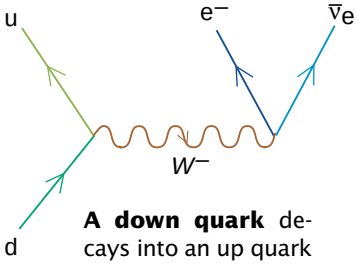
If particles could leave traces, their interactions might look like Feynman diagrams. Each line in such a diagram describes the path of a particle; when a particle breaks into two, its line divides as well. A mathematical expression is associated with each line and vertex in a Feynman diagram. The product of these expressions gives the probability that the depicted interaction occurs. Thus, Feynman diagrams are invaluable as calculation tools.



An up quark and an antiup quark ($\bar{u}$) combine to produce a Z^0 particle, which decays into an electron and a positron (e^+). This neutral current interaction is the process by which the proton-antiproton collider at CERN produced the first Z^0 particles.

A down quark and a positron interact by exchanging a W^- particle in a charged current interaction. Thereby the down quark turns into an up quark, giving off an antineutrino ($\bar{\nu}_e$). Note that the electric charge entering a vertex equals the charge going out.





A down quark decays into an up quark by emitting a W^-. The latter breaks up into an electron and an antineutrino. This diagram, describing a beta decay, is mathematically equivalent to the previous one; the incoming positron has simply been replaced by an outgoing electron. Thus, "decays" in particle physics are synonymous with "interactions."

symmetry is the existence of flavor-changing neutral currents. For example, supersymmetric particles would provide a pathway for bottom quarks to change into strange quarks. In fact, the FCNCs might be so large that they would have to be suppressed somehow.

The FCNCs mediated by SUSY particles can be reduced if the partners in a supersymmetric pair have rather similar masses. The similarity implies that SUSY particles have low masses, like those already known. But because experimenters have seen none of these particles in accelerators, their masses must actually be much heavier. They are supposed to range from 100 GeV to 10 TeV (1 TeV is a trillion electron volts). These contradictory requirements for the masses have put most versions of supersymmetry in trouble.

A more straightforward way in which the Standard Model may be extended is by additional quarks. Physicists have speculated on the possibility of a fourth family of quarks for years [see "Beyond Truth and Beauty: A Fourth Family of Particles," by David B. Cline; SCIENTIFIC AMERICAN, August 1988].

Because grand unification suggests that the quark families are also related to leptons, electrons and neutrinos are cousins of the up and down. If physicists were to find an additional, fourth neutrino, it would indicate the presence of a fourth quark family. Data taken at the Large Electron Positron collider at CERN indicate that only three light neutrinos exist. Still, there may well be a fourth, massive neutrino.

The massive quark family that would come along with a massive neutrino would almost certainly induce flavor-changing processes. As noted, GIM mechanisms, which cancel FCNCs for low-mass quarks, would not work so well with the heavier quarks. Flavor-changing events would take place most often in reactions involving the third family, into which the fourth family would preferentially decay.

Another theory has recently been put forward by Weinberg and Lawrence J. Hall of the University of California at Berkeley, as well as by some other theorists. They argue that there is no theoretical constraint on the number of Higgs particles that exist in nature. Whereas the Standard Model requires only one Higgs, it does not rule out the presence of many.

These extra Higgs particles could exist even at the relatively low mass of 100 GeV. Although hard to detect in current accelerators—because they are not very reactive—the particles would almost certainly mediate flavor-changing decays. Such decays would be most pronounced for bottom, and possibly top, quarks.

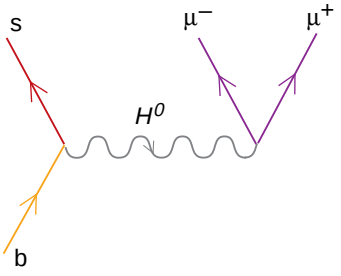
Another theory, known by the name of technicolor, suggests that the Higgs particle is a composite of two higher mass particles. This postulate allows the Higgs mechanism—by which the W and Z particles get their mass—to have a more natural structure. The technicolor particles have masses likely above a trillion electron volts. Technicolor particles also tend to generate rather large FCNCs, which are currently unapparent. Refined versions of the theory—called running technicolor or walking technicolor—manage to reduce, but not eliminate, flavor-changing currents.

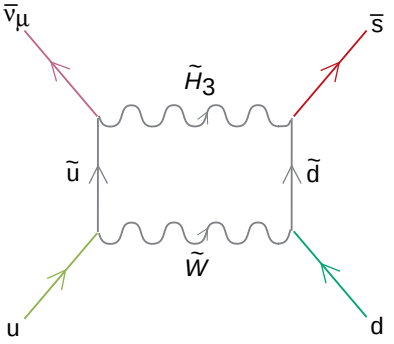
Thus, theorists predict a plethora of particles beyond the Standard Model that could give rise to FCNCs. Experimenters have looked for such currents for some 30 years now, reaching ever increasing levels of sensitivity.

Preliminary searches for neutral currents began, as mentioned, in the early 1960s. We used a kaon beam at Lawrence Berkeley Laboratory for the first definitive search. A kaon has one strange quark coupled with an antiup or antidown quark. Alternatively, it may have an antistrange quark coupled with an up or down. Kaons belong to a class of composite particles, each made of a quark and an antiquark, that are called mesons. Whereas quarks do not exist freely in nature, mesons do—although they are often unstable. Hence, experiments often begin with a meson beam.

If the strange quark in a kaon were to decay into a down, the kaon would break up into a pion—a meson that combines a down with an antiup (or up with an antidown) quark. The decaying kaon would emit as well a neutrino and

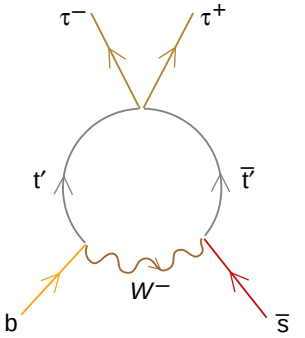
A bottom quark changes into a strange quark by emitting a muon and an antimuon (μ^+). This hypothetical decay, requiring a flavor-changing neutral current, could be mediated by an exotic Higgs particle (H^0). The UA1 detector was the first to search for the decay.





An up and a down quark may interact by exchanging supersymmetric particles ($\tilde{W}$ and $\tilde{H}_3$) to become an antistrange quark ($\bar{s}$) and a muon antineutrino ($\bar{\nu}_\mu$). Because a proton contains a down quark and two up quarks, it might decay in this manner. A search for this decay is planned at the Gran Sasso Laboratory in Italy.

A bottom quark and an antistrange quark, which make up a B meson, decay via a penguin diagram into a tau lepton and an antitau (τ^+). The two particles in the loop are a hypothetical heavy quark (t′) and its antiquark ($\bar{t}'$). This unobserved flavor-changing process may also proceed via a top quark and an antitop quark or through exotic new particles.



an antineutrino. A pion is all too common; it is made in many nuclear processes. But the two neutrinos that would come along with it are a distinctive signal of the flavor-changing process.

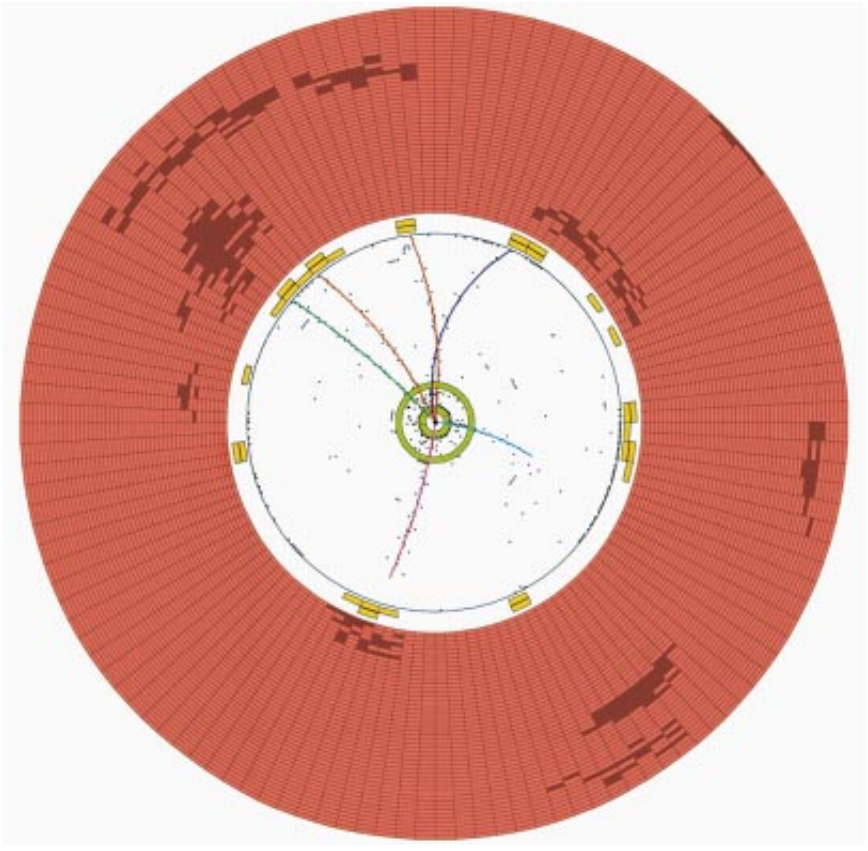
Observing the decay in an experiment is not so easy. The trace of a neutrino, for example, is never seen in a detector. Nowadays the extreme sensitivity of this search [*see bottom illustration on page 42*] has placed severe constraints on extensions of the Standard Model.

The next quark, the charm—a heavy relative of the strange—was until recently thought to be not a sensitive gauge of exotic physics. This was because it decays relatively fast, by Standard Model processes. Now we think it is interesting, for a different reason. The charm is weakly coupled to the top quark; thus, the top could decay into the charm, emitting neutrinos of very high energy. Interactions of neutrinos with charm quarks could also signal FCNCs. The latter processes could possibly be tested in future Fermilab experiments involving neutrino beams.

The most likely particle to reveal flavor-changing neutral currents is the bottom quark. Being much heavier than the strange or the charm, the bottom quark couples better with the heavy particles that are predicted by extensions of the Standard Model. Furthermore, bottom quarks are found in *B* mesons, which have a relatively long lifetime of 10^{-12} second—100 times longer than expected. The stability of *B* mesons allows experimenters to produce them in large numbers and in beams of high energy.

The bottom quark can decay in several ways via FCNCs. Any one of these decays could signal novel physics beyond the Standard Model. Besides being able to make *B* meson beams, we can now also use some extremely sensitive detectors. The *B* meson travels only a tenth of a millimeter before it decays. The latest detectors contain silicon strips in which the mesons and other particles leave tracks of electron charge. Even the very short tracks are clearly visible.

In one process, the bottom quark could decay to a strange quark by emitting an unknown object, possibly a supersymmetric particle or an exotic Higgs. The latter decays further, into a lepton and antilepton pair.



PENGUIN DECAY of a *B* meson was observed in June 1993 at the Cornell Electron Storage Ring. The collider produced a pair of *B* mesons. One decayed conventionally into a positive kaon (*green*), a negative pion (*purple*) and a photon, seen as a dark patch (*bottom right*). The other decayed via a flavor-changing neutral current, the end products of which are a negative kaon (*blue*), two positive pions (*red*), a negative pion (*pink*) and a photon (*patch at top left*). The flavor-changing decay may signal an exotic particle not within the Standard Model.

The most sensitive search to date for this decay was carried out by our group, in the unimaginatively dubbed UA1 (Underground Area 1) detector, at the CERN proton-antiproton collider. (In 1983 the UA1 collaboration reported the first observation of *W* and *Z* particles.) We looked for a muon-antimuon pair with a combined energy of more than 4 GeV. We found that fewer than five decays in 100,000 were flavor changing. The result was used to restrict the masses of technicolor and Higgs particles. If the particles interact as strongly as theorists believe them to, their masses must be less than 400 GeV.

In a different decay process, the bottom breaks down again to a strange quark, but by emitting a photon. The decay proceeds via a penguin diagram. In practice, the decaying bottom quark is contained in a *B* meson; the latter decays to an excited state of a kaon and gives off a photon.

In late 1993 such a decay was seen at the Cornell electron-positron storage ring. Only a few such events have been detected so far. Calculating the likelihood of this process is quite difficult. In particular, its presence could be signaling an exotic particle or an interaction involving a top quark. We know for sure only that it signals a penguin process. Until the decays take place frequently enough to be studied systematically, physicists cannot decide exactly which particles are mediating the penguin. At present, the finding serves to whet the appetite.

Another interaction—free of many of the theoretical uncertainties that plague the former—is one in which the *B* meson decays to any particle containing a strange quark, giving off a photon. The process includes the earlier one as a small component but is easier to calculate. Currently experimental limits have been placed on this process from the Cornell experiment. Of every 10,000 *B* meson decays, fewer than five change flavor.

There is another exciting possibility for the decay of a bottom quark. It involves a flavor-changing neutral current in which a *B* meson decays, not to another quark but to a pair of leptons. In particular, the *B* could decay to a tau and an antitau. Grand unification puts the tau lepton in the same family as bottom quarks. Thus, this decay involves only the third family. Besides, it requires a flavor-changing neutral cur-

LARGE HADRON COLLIDER is mocked-up at the tunnel currently housing the Large Electron Positron collider at CERN. The cylindrical magnets for the LHC have been placed on top of existing magnets. The LHC is planned to start operating in 2003 and will shoot two beams of protons at each other at higher energies than have ever been achieved. The resulting collisions will, it is hoped, create the Higgs particle and provide evidence for particles beyond the Standard Model.

rent. If the decay is relatively profuse, it would point to the existence of supersymmetric particles.

Detecting this decay is a major challenge to experimental particle physics. At a recent meeting in Snowmass, Colo., a few of us initiated a study of schemes for its observation. To this end, we are conducting a series of computer simulations at the University of California at Los Angeles.

One approach is to detect the muons into which the tau lepton decays. A key detector in this search is the just approved Compact Muon Solenoid. It is to be used at the Large Hadron Collider (LHC) at CERN. Our group is part of a collaboration that designed and, we hope, will participate in building the detector. The current head of this experiment is Michel Della Negra of CERN.

In addition to detection schemes researchers also require intense sources of *B* particles. One such source might be derived from the proton-antiproton beams at Fermilab. When the two beams collide, they generate a profusion of particles, including between 10^9 to 10^{10} *B* mesons. Two "*B* factories" are being planned as well, at the Stanford Linear Accelerator Center and at the National Laboratory for High Energy Physics (better known as KEK) in Japan. These projects should each produce about 10^8 *B* mesons.

Colliders to be built in the future will also be important for such searches. The European Union is going ahead with the LHC. This collider will smash together, head-on, two proton beams, each with energies of 7 TeV. If all goes as planned, the LHC will turn on before the year 2003. It will create some 10^{12} *B* mesons in colliding beams. Another possible means of detecting *B* decays at the LHC is the super fixed target experiment. If a part of the main beam is extracted and made to hit a stationary target, up to 10^{11} *B* mesons could be manufactured.

Many teams from the U.S. are now planning to work at the LHC. A subpanel of the High Energy Physics Advisory Panel, chaired by Sidney D. Drell of the Stanford Linear Accelerator, recently emphasized to the U.S. Department of Energy the need to support such participation. Fortunately for those of us at U.C.L.A., our early involvement in the Compact Muon Solenoid guarantees our place in the LHC.

The discovery of the top quark gives physicists a more accurate tool in evaluating decays of the bottom quark. Now that the mass of the top is known, theorists can calculate the frequency of penguin processes involving top quarks. Knowing the top's contribution, they can more precisely gauge which FCNCs signal exotic particles.

The top quark could also decay in exotic ways that signal unusual physics. For instance, it might decay to a charm and two neutrinos, a decay mediated by technicolor or multiple Higgs particles. The high mass of the top—174 GeV—might be part of a general pattern, indicating that exotic particles are even heavier than theorists had anticipated. They could range from hundreds of GeV to 1 TeV.

The observations of flavor-changing decays at Cornell and the limits on exotic particles from UA1 have put scientists in a new era of searches for phenomena beyond the Standard Model. With the profuse sources of *B* mesons experimenters will have in the near future, and information about top quarks, they can consolidate the early sightings of flavor-changing processes—and tease out the implications.

The story of flavor-changing neutral currents illustrates the role that "null" experiments—those that see nothing—have played in guiding the development of particle physics. We hope the 30 years of arduous searches will be rewarded in the not too distant future with more discoveries. Even before the Large Hadron Collider comes on line, physicists may be able to peel partially yet another layer from the elementary-particle onion.

FURTHER READING

A TOUR OF THE SUBATOMIC ZOO. C. Schwarz, AIP Press, 1993.

DISCOVERY OF WEAK NEUTRAL CURRENTS: THE WEAK INTERACTION BEFORE AND AFTER. Edited by A. K. Mann and D. B. Cline. AIP Press, 1994.

THIRTY YEARS OF WEAK NEUTRAL CURRENTS. D. B. Cline in *Comments in Nuclear and Particle Physics,* Vol. 21, No. 4, pages 193–222; March 1994.

The Aluminum Beverage Can

Produced by the hundreds of millions every day, the modern can—robust enough to support the weight of an average adult—is a tribute to precision design and engineering

by William F. Hosford and John L. Duncan

Makers of beer and soft-drink containers in the U.S. produce 300 million aluminum beverage cans a day, 100 billion of them every year. The industry's output, the equivalent of one can per American per day, outstrips even the production of nails and paper clips. If asked whether the beverage can requires any more special care in its manufacture than do those other homey objects, most of us would probably answer negatively. In fact, manufacturers of aluminum cans exercise the same attention and precision as do makers of the metal in an aircraft wing. The engineers who press the design of cans toward perfection apply the same analytical methods used for space vehicles.

As a result of these efforts, today's can weighs about 0.48 ounce, down from about 0.66 ounce in the 1960s, when such containers were first constructed. The standard American aluminum can, which holds 12 ounces of liquid, is not only light in weight and rugged but is also about the same height and diameter as the traditional drinking tumbler. Such a can, whose wall surfaces are thinner than two pages from this magazine, withstands more than 90 pounds of pressure per square inch—three times the pressure in an automobile tire.

Yet the can industry is not standing pat on its achievement. Strong economic incentives motivate it toward further improvements. Engineers are seeking ways to maintain the can's performance while continuing to trim the amount of material needed. Reducing the can's mass by 1 percent will save approximately $20 million a year in aluminum (and make still easier and even less meaningful the macho gesture of crushing an empty can with a bare hand).

Aside from the savings it yields, the modern manufacturing process imparts a highly reflective surface to the can's exterior, which acts as a superb base for decorative printing. This attribute adds to the enthusiasm for the aluminum can among those who market beverages. Indeed, that industry consumes about a fifth of all aluminum used in the U.S. Consequently, beverage cans have emerged as the single most important market for aluminum. Until 1985, most cans held beer, but now two thirds of them store nonalcoholic drinks.

The aluminum beverage can is a direct descendant of the steel can. The first of these vessels appeared in 1935, marketed by Kreuger Brewing Company, then in Richmond, Va. Similar to food cans, this early beverage container comprised three pieces of steel: a rolled and seamed cylinder and two end pieces. Some steel cans even had conical tops that were sealed by bottle caps. During World War II, the government shipped great quantities of beer in steel cans to servicemen overseas. After the war, much of the production reverted to bottles. But veterans retained a fondness for canned beer, so manufacturers did not completely abandon the technology even though the three-piece cans were more expensive to produce than the bottles.

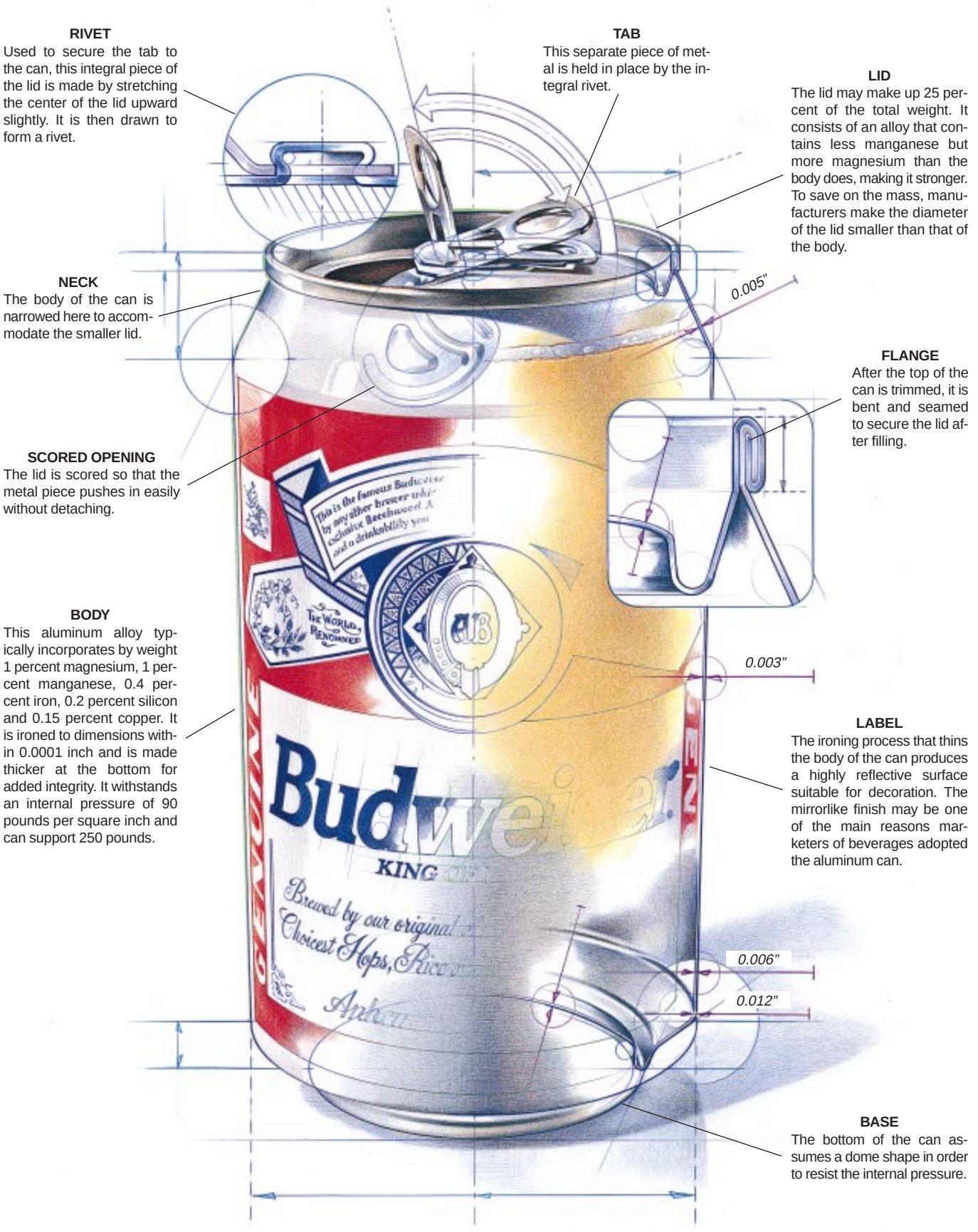
The first aluminum beverage can went on the market in 1958. Developed by Adolph Coors Company in Golden, Colo., and introduced to the public by the Hawaiian brewery Primo, it was made from two pieces of aluminum. To produce such cans, Coors employed a so-called impact-extrusion process. The method begins with a circular slug that has a diameter equal to that of the can. A punch driven into the slug forces material to flow backward around it, forming the can. The process thus made the side walls and the bottom from one piece. The top was added after filling.

This early technique proved inadequate for mass manufacturing. Production was slow, and tooling problems plagued the process. Moreover, the resulting product could hold only seven ounces and was not efficient structurally: the base could not be made thinner than 0.03 inch, which was much thicker than it needed to be to withstand the internal forces.

Nevertheless, the popularity of the product encouraged Coors and other companies to look for a better way to make the cans. A few years later Reynolds Metals pioneered the contemporary method of production, fabricating the first commercial 12-ounce aluminum can in 1963. Coors, in conjunction with Kaiser Aluminum & Chemical Corporation, soon followed. But pressure from large can companies, which also purchased steel from Kaiser for three-piece cans, is said to have obliged Kaiser to withdraw temporarily from aluminum-can development. Apparently, these steel-can makers feared the competition of a new breed of container. Hamms Brewery in St. Paul, Minn., began to sell beer in 12-ounce aluminum cans in 1964. By 1967 Coca-Cola and PepsiCo were using these cans.

Today aluminum has virtually displaced steel in all beverage containers. The production of steel three-piece cans, which are now rarely made, reached its peak of 30 billion cans in 1973. The number of two-piece steel cans topped out at 10 billion in the late 1970s. This design now accounts for less than 1 percent of the cans in the U.S. market (they

WILLIAM F. HOSFORD and JOHN L. DUNCAN have been active in research on sheet-metal forming for more than 30 years and act as consultants to aluminum producers. Hosford is professor of materials science and engineering at the University of Michigan. He received his doctorate in metallurgical engineering from the Massachusetts Institute of Technology and has written books on metal forming and the plasticity of materials. Duncan, who received his Ph.D. in mechanical engineering from the University of Manchester in England, is professor of mechanical engineering at the University of Auckland in New Zealand. Like Hosford, Duncan has written a textbook on the forming of sheet metal.



ANATOMY OF MODERN BEVERAGE CAN reveals the dimensions that design and engineering must achieve on a daily basis. The goal of can makers is to reduce the amount of aluminum needed without sacrificing structural integrity. A can now weighs about 0.48 ounce; the industry hopes to reduce that weight by about 20 percent.

STEPS IN CAN MANUFACTURE begin with an aluminum alloy sheet. Blanks 5.5 inches in diameter are cut from the sheet; a punch draws the circle to form a 3.5-inch-diameter cup. A second machine then redraws the blank, irons the walls and gives the base its dome—all in approximately one fifth of a second. These procedures give the can wall its final dimen-

are, however, more popular in Europe).

The process that Reynolds initiated is known as two-piece drawing and wall ironing. Aluminum producers begin with a molten alloy, composed mostly of aluminum but also containing small amounts of magnesium, manganese, iron, silicon and copper. The alloy is cast into ingots. Rolling mills then flatten the alloy into sheets.

The first step in can making is cutting circular blanks, 5.5 inches in diameter. Obviously, cutting circles from a sheet produces scrap. The theoretical loss for close-packed circles is 9 percent; in practice, the loss amounts to 12 to 14 percent. To reduce this waste, sheets are made wide enough to incorporate 14 cups laid out in two staggered rows. Each blank is drawn into a 3.5-inch-diameter cup.

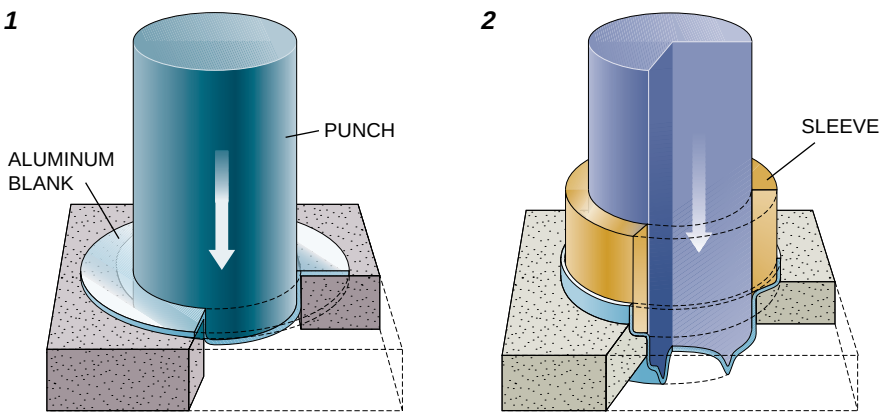
The next three forming operations for the can body are done in one continuous punch stroke by a second machine—in about one fifth of a second. First, the cup is redrawn to a final inside diameter of about 2.6 inches, which increases the height from 1.3 to 2.25 inches. Then, a sequence of three ironing operations thins and stretches the walls, so that the body reaches a height of about five inches. In the last step, the punch presses the base of the can body against a metal dome, giving the bottom of the can its inward bulge. This curve behaves like the arch of a bridge in that it helps to prevent the bottom from bulging out under pressure. For added integrity, the base of the can and the bottom of the side walls are made thicker than any other part of the can body.

Because the alloy does not have the same properties in all directions, the can body emerges from the forming operations with walls whose top edges are wavy, or "eared." To ensure a flat top, machinery must trim about a quarter inch from the top. After trimming, the cup goes through a number of high-speed operations, including washing, printing and lacquering. Finally, the can is automatically checked for cracks and pinholes. Typically, about one can in 50,000 is defective.

Ironing is perhaps the most critical operation in making the body of the can. The precisely dimensioned punch holds and pushes the cup through two or three carbide ironing rings. To thin and elongate the can, the punch must move faster than the metal does in the ironing zone. The clearance between the punch and each ring is less than the thickness of the metal. The friction generated at the punch surface assists in pushing the metal through the ironing rings. To increase this friction, the punch may be slightly roughened with a criss-cross scratch pattern (which can be seen, impressed on the inside of a can). On the exterior of the can the shearing of the surface against the ironing rings yields the desired mirror finish.

The side walls can be thinned without loss of integrity because, structurally, the can is a "pressure vessel." That is, it relies for part of its strength on the internal force exerted by carbon dioxide in beer and soft drinks or by the nitrogen that is now infused into such uncarbonated liquids as fruit juice. Indeed, most beers are pasteurized in the can, a process that exerts nearly 90 pounds per square inch on the material. Carbonated beverages in hot weather may also build up a similar pressure.

Filling introduces a different kind of stress on the can. During this stage, the can (without its lid) is pressed tightly against a seat in a filling machine. It must not buckle, either during filling and sealing or when filled cans are stacked one on another. Hence, can makers specify a minimum "column strength" of



DRAWING AND IRONING constitute the modern method of beverage can manufacture. The initial draw transforms the blank into a small cup (*1*). The cup is trans-

sions. After the "ears" at the top of the walls are trimmed, the can is cleaned, decorated and then "necked" to accommodate the smaller lid. The top is flanged to secure the lid. Once filled and seamed shut, the can is ready for sale.

about 250 pounds for an empty can body. Thin-walled structures do not easily meet such a requirement. The slightest eccentricity of the load—even a dent in the can wall—causes a catastrophic collapse. This crushing can be demonstrated by standing (carefully) on an upright, empty can. Manufacturers avoid failures by using machines that hold the cans precisely.

The second piece of the can, the lid, must be stiffer than the body. That is because its flat geometry is inherently less robust than a curved shape (dams, for instance, bow inward, presenting a convex surface to the waters they restrain). Can makers strengthen the lid by constructing it from an alloy that has less manganese and more magnesium than that of the body. They also make the lid thicker than the walls. Indeed, the lid constitutes about one fourth the total weight of the can. To save on the mass, can makers decrease the diameter of the lid so that it is smaller than the diameter of the cylinder. Then they "neck down" the top part of the cylindrical wall, from 2.6 to 2.1 inches, to accommodate the lid. An ingenious integral rivet connects the tab to the lid. The lid is scored so that the can opens easily, but the piece of metal that is pushed in remains connected.

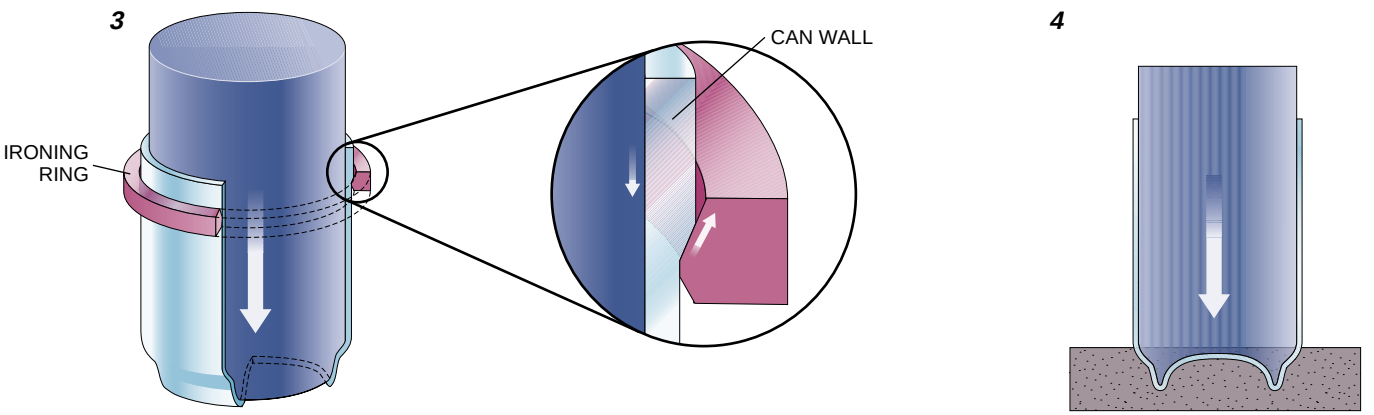
In addition to clever design, making billions of cans a year demands reliable production machinery. It has been said that in order to prove himself, an apprentice Swiss watchmaker was not required to make a watch but rather to make the tools to do so. That sentiment applies to can manufacturing. As one production manager remarked, "If at the end of a bad day, you are a half million cans short, someone is sure to notice." A contemporary set of ironing dies can produce 250,000 cans before they require regrinding. That quantity is equivalent to more than 20 miles of aluminum stretched to tolerances of 0.0001 inch. Die rings are replaced as soon as their dimensions fall out of specification, which occurs sometimes more than once a day.

Much of the success behind the consistent and precise production lies in the strong yet formable alloy sheet. The metallurgical properties responsible for the performance of modern can sheet have been proprietary and therefore not well known. Only within the past decade has that situation changed. Through the efforts of Harish D. Merchant of Gould Electronics in Eastlake, Ohio, James G. Morris of the University of Kentucky and others, scientific papers on the metallurgy of can sheet have become more widely published.

We now know that three basic factors increase the strength of aluminum. We have already mentioned one of them: manganese and magnesium dissolved into the material. These atoms displace some of the aluminum ones in the substance. Because they are slightly different in size, the manganese and magnesium atoms distort the crystal lattice. The distortions resist deformation, thus adding strength to the sheet.

The second contribution comes from the presence of so-called intermetallic particles. Such particles, which form during the processing of the sheet, consist of a combination of different metals in the alloy (mostly iron and manganese). They tend to be harder than the alloy itself, thus supplying strength.

Perhaps the most important contribution to sheet strength, however, is the work hardening that occurs when the sheets are cold-rolled (flattened at room temperature). During this shaping, dislocations, or imperfections, in the lattice materialize. As the metal deforms, the dislocations move about and increase in number. Eventually they become entangled with one another, making further deformation more difficult.



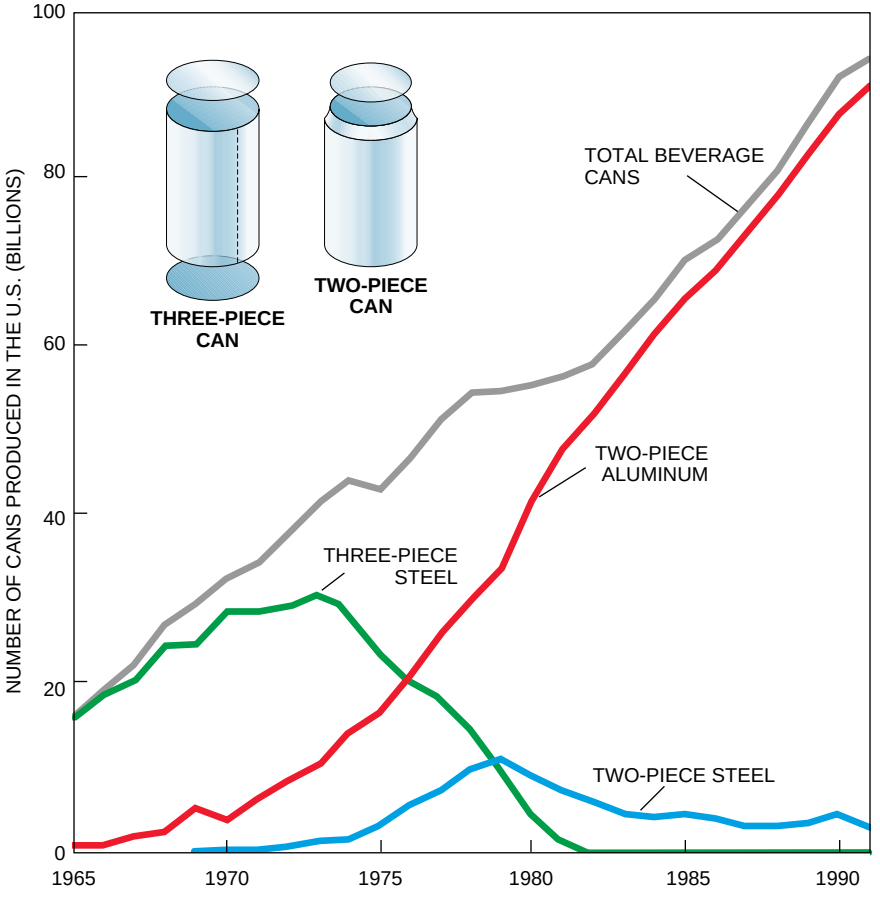
ferred to a second punch, which redraws the can; the sleeve holds the can in place to prevent wrinkling (*2*). The punch pushes the can past ironing rings, which thin the walls (*3*). Finally, the bottom is shaped against a metal dome (*4*).

Unfortunately, this work hardening dramatically reduces the ability of the material to stretch. Tensile tests indicate that the elongation capacity drops from 30 percent to about 2 or 3 percent. Conventional wisdom had it that sheets can be formed only if the material has a high tensile elongation. Certainly in the automotive industry, body parts are formed from fully annealed sheets that can elongate more than 40 percent. This philosophy guided the early attempts to make two-piece aluminum cans. Researchers concentrated on annealed or partially work-hardened sheets, which sacrificed strength for ductility.

The understanding of formability received a major boost from studies in the 1960s by Stuart P. Keeler and Walter A. Backofen of the Massachusetts Institute of Technology and Zdzislaw Marciniak of the Technical University in Warsaw, among others. Looking at the behavior of various sheet metals, they considered more than just the behavior under tension applied in one direction (as is done in the tension test). They also looked at what happens when tension is applied simultaneously in two directions. They showed that a small window of strains exists that permits forming without structural failure. Although work hardening greatly reduces the size of this window, a small slit nonetheless remains open—enough to permit the doming of the base and drawing and redrawing of the side walls.

The crucial advance that made the aluminum can economical, however, came from Linton D. Bylund of Reynolds. He realized that cans could be made from a fully work-hardened sheet using a carefully designed process that specified the placement of the ironing rings, the shape of the punch and dies, and many other parameters. The strong, fully work-hardened sheet made it possible to use sheet that was thinner, saving enough weight to make the cans economically competitive.

Nowhere is the technique of forming work-hardened sheet more apparent than it is in the cleverly designed rivet that holds the tab on the can lid. The rivet is an integral piece of the lid. To make it, the center of the lid must be stretched by bulging it upward a bit. This "extra" material is drawn to form a rivet and then flattened to secure the tab (which is a separate piece of metal).



ANNUAL BEVERAGE CAN PRODUCTION in the U.S. has increased by several billion over the past few years. The two-piece aluminum can overwhelmingly dominates the market; steel cans constitute less than 1 percent. Three-piece steel cans, which are now rarely made, reached their peak production in the mid-1970s.

Besides making the can sheet stronger, manufacturers also sought to reduce the amount of aluminum needed by controlling the waviness, or earing, which as we have seen takes place at the top of the can after ironing. The effect derives from the crystallographic texture of the aluminum sheet, that is, the orientation of its crystal structure. Hence, earing is inevitable to some extent. Hans-Joachim Bunge of the Technical University in Clausthal, Germany, and Ryong-Joon Roe of Du Pont and others have developed x-ray diffraction techniques to describe qualitatively the textures that cause earing. Laboratory technicians prepare specimens by grinding away layers of the sheet to expose material at different depths. X-ray diffraction coupled with elegant analytical techniques automatically produces three-dimensional diagrams that reveal the preferred orientation of crystals as a function of depth in the sheet.

Such diagnostic approaches have enabled aluminum companies to produce sheet that yields much smaller ears. Metallurgists balance the two predominant crystallographic textures that exist in the aluminum. One kind of texture arises during annealing of the alloy after the alloy is hot-rolled from ingots. It causes four ears to appear every 90 degrees (at 0, 90, 180 and 270 degrees) around the circumference of the can. The second kind of texture results from cold-rolling the sheet, which produces an ear at 45, 135, 225 and 315 degrees. Proper control of annealing and rolling can lead to a combination of the two textures such that ears caused by one fill the valleys caused by the other. The result is eight very low ears. The maximum height of an ear is often less than 1 percent of the height of the cup.

Consistent processing of metal and careful design have now made each part of the can about as strong as any other. It is not unusual to find cans in which the opening on the lid fractures, and the bottom dome and lid bulge at nearly the same pressure, within the range of 100 to 115 pounds per square inch.

Despite the success of current design and manufacture, can makers are still searching for refinements. Much of the investigation focuses on ways to use aluminum more efficiently, because the metal represents half the cost of the can. One possibility for saving would be to cast the molten alloy into thin slabs rather than into thick ingots, as is currently done. A typical ingot may be 30 inches thick, which is rolled down

by a factor of 2,500 to 0.011 or 0.012 inch. So much rolling requires expensive capital equipment—furnaces and rolling mills—and consumes a lot of energy.

It is possible to cast aluminum continuously into slabs that are an inch thick or less. These thin slabs would require much less rolling to reach the desired final sheet thickness. Continuous casting is used for some soft aluminum alloys—for example, aluminum foil is made from material cast to a thickness of 0.1 inch.

Unfortunately, production of satisfactory can stock from thin slabs thwarts the metallurgists. The faster cooling and decreased rolling inherent in continuous casting do not yield the desired metallurgical structure. Two main problems arise. First, crystallographic texture cannot be properly controlled to prevent large ears. Second, the faster cooling rate produces severe difficulties in ironing the can walls.

These ironing problems develop because of the nature of the intermetallic particles that form when the molten alloy solidifies. Intermetallic particles that develop during solidification are much larger than those that originate during processing (which as we have seen impart strength to the sheet). Because of their size, they play a key role in ironing. During this procedure, aluminum tends to adhere to the ironing rings. Ordinarily, the intermetallic particles, which are about five microns in size, act like very fine sandpaper and polish the ironing rings. The faster cooling rates of continuous casting, however, produce intermetallic particles that are much smaller (about one micron). At this size, the particles are not very effective in removing aluminum that sticks to the ironing rings. As a result, aluminum builds up on the rings and eventually causes unsightly scoring on the can walls. The problem of achieving thin slabs with the desired intermetallic particles may yet be solved, perhaps by altering the composition of the alloy or by shifting the rate of solidification from the material's molten state.

EASY-OPENING LIDS were introduced on three-piece steel cans in 1961. The original caption reads: "Housewives of ancient Greece and the space age compare containers for the kitchen at the press debut of the new canning innovation by the Can-Top Machinery Corp., Bala-Cynwyd, Pa."

The control of casting epitomizes a recurrent feature of the whole can story: one behavior is carefully traded off against another, from the control of earing and ironability to economical sheet production, from can weight to structural integrity. Yet one cost element eludes an easy balance: the energy needed to make cans. Most of this outlay lies in the aluminum itself. Taking into account inefficiencies in electricity distribution and smelting, industry experts estimate that 2.3 megajoules of energy is needed to produce the aluminum in one can. This value is equal to about the amount of energy expended to keep a 100-watt bulb lit for six hours, or about 1.7 percent of the energy of a gallon of gasoline. Although small, it represents the major expenditure of a can.

One way to reduce this expense is through recycling, which can save up to 95 percent of the energy cost. Indeed, more than 63 percent of aluminum cans are now returned for remelting. Recycling also has an important part within the aluminum mill. For every ton of can bodies made, a ton of scrap metal is produced. This scrap is remelted and thus injected back into the manufacturing cycle. Developing simpler ways of producing can sheet and finding stronger materials that can lead to lighter cans should save more money and energy.

Meeting these goals presents a great challenge. Existing cans already use a highly strengthened, well-controlled sheet. Their shape is finely engineered for structural strength and minimum weight. And with little tool wear, the production machinery in a single plant is capable of making many millions of cans a day with few defects. The rewards of even small improvements, however, are quite substantial. The demand for aluminum beverage cans continues to grow everywhere in the world; their production increases by several billion every year. The success of the can is an industrial lesson about what can be achieved when scientific and engineering skills are combined with human perseverance.

FURTHER READING

A GOLDEN RESOURCE. Harold Sohn and Karen Kreig Clark. Ball Corporation, 1987.

FROM MONOPOLY TO COMPETITION: THE TRANSFORMATIONS OF ALCOA, 1888–1986. George David Smith. Cambridge University Press, 1988.

THE MECHANICS OF SHEET METAL FORMING. Z. Marciniak and J. L. Duncan. Edward Arnold, 1992.

ALUMINUM ALLOYS FOR PACKAGING. Edited by J. G. Morris, H. D. Merchant, E. J. Westerman and P. L. Morris. Minerals, Metals and Materials Society, Warrendale, Pa., 1993.

METAL FORMING: MECHANICS AND METALLURGY. William F. Hosford and Robert M. Caddell. Prentice Hall, 1993.

The Machinery of Cell Crawling

When a cell crawls, part of its fluid cytoplasm briefly turns rigid. This transformation depends on the orderly assembly and disassembly of a protein scaffold

by Thomas P. Stossel

People are often surprised, even alarmed, to learn that many of their cells crawl around inside them. Yet cell crawling is essential to our survival. Without it, our wounds would not heal; blood would not clot to seal off cuts; the immune system could not fight infections. Unfortunately, crawling contributes to some disease processes, too, such as destructive inflammation and the formation of atherosclerotic plaques in blood vessels. Cancer cells crawl to spread themselves throughout the body: were cancer just a matter of uncontrolled cell growth, all tumors would be amenable to surgical removal.

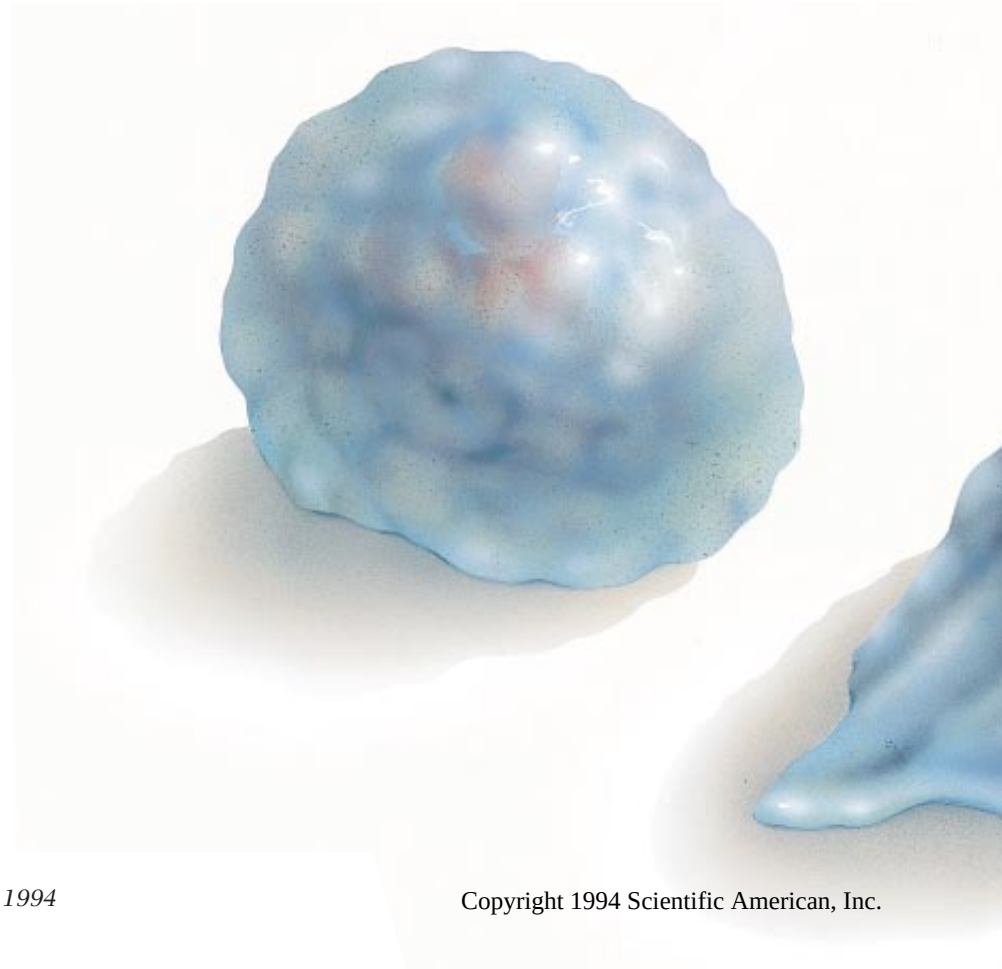
The observation of cells crawling has suggested compelling ideas about the crawling mechanism. In 1786 the Danish biologist Otto F. Müller described a crawling cell as a "clear gelatinous body from which extends a glassy spike." The term "gelatinous" was inspired by the Latin verb *gelare,* meaning "to freeze." This notion of a mechanical state change in the cell—a "sol-gel transformation," as we now call it—has been very useful for picturing the mechanism of cell crawling and for isolating the molecular components of the machinery.

It even points the way toward potential medical treatments for several kinds of illness. Infections and cancer would clearly number among these afflictions, but so, too, might cystic fibrosis.

Cells in healing wounds and cancer cells crawl relatively slowly, at rates of 0.1 to 1 micron per hour. In contrast, cells involved in body defenses against infection and hemorrhage move much faster. To fight infection, a human being produces daily more than 100 billion of the white blood cells called neutrophils. Neutrophils originate in the bone marrow, creep out of it to cruise through the bloodstream for a few hours, then crawl out of the capillaries and into other tissues. At rates of up to 30 microns per minute, these migrating cells search for and ingest microorganisms infesting the skin, airways and gastrointestinal tracts. A neutrophil will move several millimeters in this way. In fact, the aggregate distance actively traveled every day by all the neutrophils in the human body would circle the earth twice.

The cells called platelets do not locomote, but they do change their appearance through rapid crawling movements to stop bleeding. When platelets are circulating in the blood, they are tiny discoid objects. At the sites of trauma, however, they quickly spread into shapes that resemble spiny pancakes to plug leaks in injured blood vessels.

As seen through an optical microscope, cell crawling involves extensions and contractions of the cell's outer rim, or cortex. In contrast with deeper areas of the cell, which are dotted with vari-



THOMAS P. STOSSEL is the American Cancer Society Professor of Medicine at Harvard Medical School and director of the division of experimental medicine at Brigham and Women's Hospital in Boston, where he is also a senior physician in the division of hematology and oncology. A graduate of Princeton University, he received his M.D. degree from Harvard in 1967. He serves on the advisory boards of the biotechnology firms Biogen and Protein Engineering Co. and is also a member of the research council of the American Cancer Society. Stossel's research focuses on how cells crawl in the human body during immune responses and cancer metastasis. He has also written about scientific communication and the role of research in medicine.

ous subcellular organelles, the cortex appears clear and homogeneous.

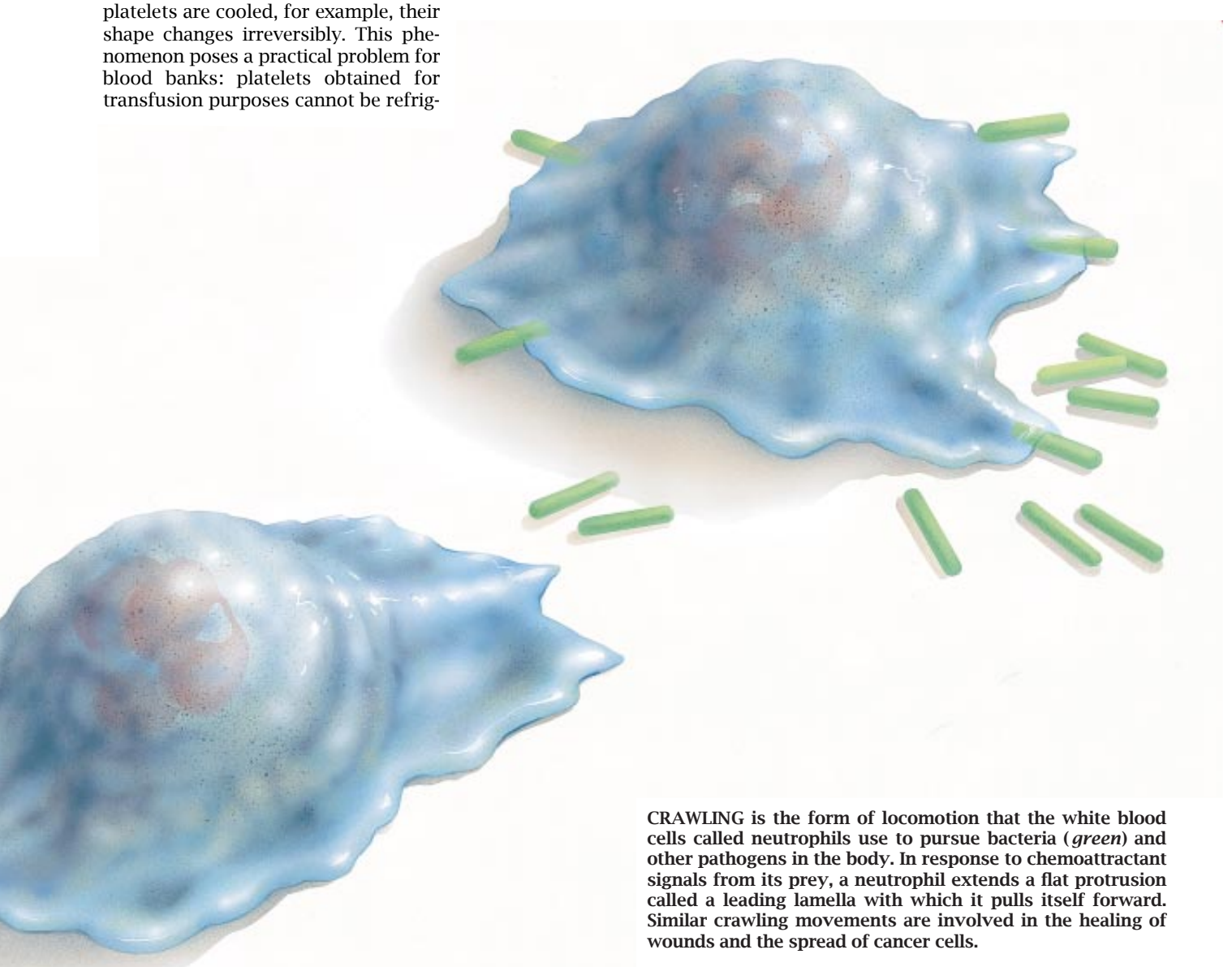
Cells crawl in response to external instructions. White blood cells follow trails of chemoattractants, diverse molecules derived from microorganisms or damaged tissues. Growth factors that trigger cell division can also induce directed cell movements. Thrombin, an enzyme modified by blood coagulation reactions, makes platelets change shape.

Most agents that inaugurate cell crawling work by first reacting with specific receptors on the outer membrane of the cells. Ligation with the receptors then elicits a sequence of molecular reactions, collectively known as signal transduction, that controls the cortical rearrangements responsible for the crawling motions. In addition, however, some other stimuli, such as low temperatures, can apparently bypass the membrane receptors and still cause these cortical changes. If platelets are cooled, for example, their shape changes irreversibly. This phenomenon poses a practical problem for blood banks: platelets obtained for transfusion purposes cannot be refrigerated to slow their degeneration and minimize bacterial growth.

As a cell begins to crawl, part of its cortex flows out to form a flat projection known as the leading lamella. Early microscopists described these lamellae as "hyaline," meaning "glassy," because of their lack of organelles. Hairlike projections called filopodia supply the excess membrane to accommodate the lamellar extensions; they are also used to pull objects back to the cell. The bottom of the lamella attaches to the underlying surface, primarily through the action of membrane-adhesion proteins. Binding between these proteins and molecules on the substrate provides a traction force that enables the cell body to pull itself forward. The lamella then detaches from the substrate and flows forward yet again. The protrusion, attachment, contraction and detachment steps are often so tightly coordinated that the cell appears to glide along, like a cloud against a mountainside.

During these movements, the cell body behaves like a sol, a liquid that flows in response to an applied stress. Yet if you were to poke the leading lamella with a microscopic needle or to try to pull it into a capillary tube, you would find that it resists deformation. Thus, the cell body is also a gel—an elastic structure that is primarily liquid but has some solid properties. The cell body deforms in response to applied stress, but it has a memory of its starting configuration and exhibits elastic recoil when the stress is removed. The ratio of this elastic deformation to the applied stress is the modulus of rigidity.

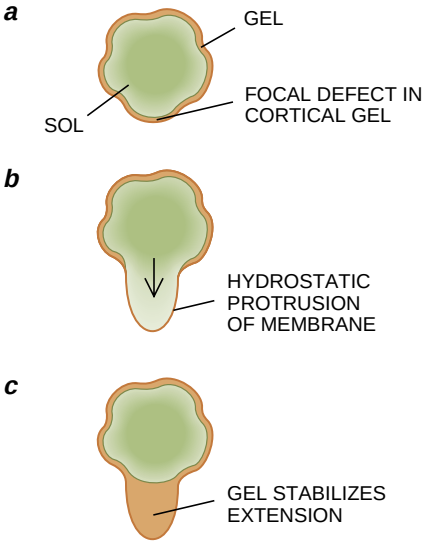
Gels also have important ionic and hydraulic properties, which include the ability to retard the flow of a solvent, much as a sponge holds water. The elastic and water-retaining properties of the cell cortex come from water-soluble polymers in the cytoplasm. These polymers also serve as scaffoldings for the imposition of contractile forces.



CRAWLING is the form of locomotion that the white blood cells called neutrophils use to pursue bacteria (*green*) and other pathogens in the body. In response to chemoattractant signals from its prey, a neutrophil extends a flat protrusion called a leading lamella with which it pulls itself forward. Similar crawling movements are involved in the healing of wounds and the spread of cancer cells.

Dennis C. H. Bray of the Medical Research Council has articulated a mechanism for cell crawling based on venerable ideas about sol-gel transformations. He envisioned the cell as a sol encased in a rim of gel. When the cell is stimulated, the gel is subject to contractile tension. The sol is not compressible, however, so nothing happens until the gel weakens at the point of initial or maximal stimulation. Hydrostatic force then causes the cell membrane to protrude smoothly at the site of the defect. This protrusion immediately fills with new gel substance and becomes the glassy lamella seen by early microscopists. In this way, the cell advances. If the defect in the gel occurs at the base of the lamella rather than at its tip, the protrusion retracts into the cell body.

This model does explain the observed behavior of the crawling cells well. The challenge for cell biologists has been to identify the molecular structure of the cortical gel and to elucidate how the cellular material transforms smoothly and rapidly between the sol and gel states in response to stimuli.



HYDROSTATIC MODEL of a cell can explain how crawling occurs. The cell body consists of a sol, or fluid, surrounded by a more rigid gel. When stimulation weakens the gel at one point (*a*), hydrostatic pressure causes the sol to push out the cell membrane (*b*). Material in the protrusion immediately turns to gel (*c*), forming a stable lamella.

When my colleagues and I began to study the molecular mechanism of cell crawling in the early 1970s, the front-running candidates for the molecular machinery were the proteins actin and myosin. Since the 1940s actin and myosin had been known as the major proteins of skeletal muscle, and in the 1960s Sadashi Hatano and Fumio Oosawa of Nagoya University had also found them in amoeboid nonmuscle cells. Other investigators showed that actin, which constitutes 10 percent of the total protein of neutrophils and 20 percent of blood platelets, concentrates in the cell cortex and in the leading lamella.

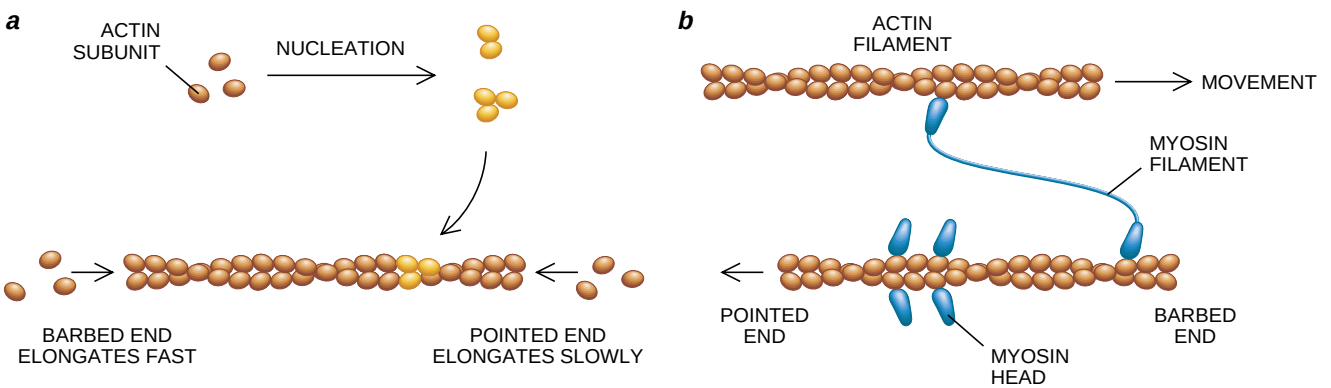
Actin had been characterized as a globular protein that polymerizes into long double helical filaments. Myosin molecules were known to be more complex: they have globular "head" domains that bind actin filaments and helical "tail" domains that self-associate to form bipolar myosin filaments. In 1963 Hugh E. Huxley of the MRC demonstrated that myosin heads bind to actin filaments at an acute angle; under the electron microscope, the filaments, adorned with myosin heads, looked like a series of arrowheads. Investigators therefore defined the two poles of the actin filament as "pointed" and "barbed."

Huxley had also been instrumental in proposing that muscle contraction results from the sliding of actin and myosin filaments in opposite directions, with the actin filaments moving in the pointed direction. The energy for this sliding motion is derived from the enzymatic decomposition of adenosine triphosphate (ATP) by the myosin heads. As Setsuro Ebashi of the University of Tokyo had proved, in muscle cells this chemical reaction is regulated by calcium ions acting on troponin and tropomyosin, two other proteins that adhere to the surface of the actin filament.

In the mid-1970s Robert S. Adelstein of the National Institutes of Health found that calcium can also control the mechanochemical activity of myosin molecules. In nonmuscle cells, calcium indirectly triggers the chemical addition of phosphate groups to the head of myosins. Once phosphorylated in this way, the myosins can exert contractile forces on actin filaments. Other enzymes remove those phosphates and inactivate the myosin. All this work suggested that cell stimulation induced the contraction of a cortical actin network by altering calcium levels and activating myosin.

My initial studies of the chemistry of cell crawling addressed the nature of the actin gel. This research began in 1974 at Harvard Medical School in collaboration with John H. Hartwig. We first discovered that if we stirred extracts of white blood cells under certain experimental conditions, large amounts of actin precipitated along with an unknown protein of high molecular weight. We purified the latter and dubbed it actin-binding protein (ABP).

At about the same time, the late Robert E. Kane of the University of Hawaii at Manoa reported that extracts of sea-urchin eggs, which were initially liquid, gelled upon standing. These gels were



ACTIN is the major protein constituent of the cortical cytoplasmic gel. When two or more actin subunits bind together and form a nucleus, other subunits rapidly attach to it and create a double-strand polar filament. Subunits attach to the "barbed" end of these filaments faster than they do to the "pointed" end (*a*). Another protein, myosin, can bind to the actin filaments and exert a pulling force that causes pairs of actin filaments to slide past one another (*b*).

filled with filamentous actin. Subsequently, we and others learned that extracts from diverse types of cells could form similar actin-rich gels. Low concentrations of ABP were associated with large amounts of actin; we therefore reasoned that ABP was responsible for the gelation of actin.

Hartwig and I demonstrated that ABP could induce abrupt increases in the elasticity of actin solutions: as little as one ABP molecule per 1,000 actin molecules in filaments could make an actin sol more rigid. No other actin-binding molecules came close to this efficiency in making actin gel. That fact was suggestive about how this gelling occurs.

If you put a collection of stiff rods (like actin filaments) into a container and shake them, entropy will drive the rods to align into parallel bundles. Inside a cell's cortex, actin filaments presumably align the same way, and various cellular proteins can cross-link these parallel filaments to give the bundles greater stability. Such parallel arrays of actin filaments confer tensile strength on filopodia. Cross-linked actin bundles can also associate with adhesion molecules to form multimolecular assemblages called adhesion plaques. Such bundles, however, are not useful for the construction of a uniform lamellar gel.

On the other hand, a protein that recruited filaments into a uniform, orthogonal, three-dimensional network would create such a gel very easily. Hartwig and I theorized that for ABP to make actin gel so efficiently, it must cause actin filaments to branch at roughly right angles. In 1981 we obtained electron micrographs of actin filaments cross-linked by ABP and found that the filaments did indeed branch at the predicted perpendicular angles.

PLATELETS, the circulating cells that enable blood to clot, also crawl, but not for locomotion. At the site of trauma, platelets use crawling movements to change from their normal discoid shape (*top*) to a flatter form (*bottom*).

As we learned more about the structure of ABP, the molecule's talent for assembling actin gels became more understandable. ABP is a very large filamentous molecule. One end of each ABP subunit binds to actin; the other end tends to associate with the like end of another ABP subunit. In between the two ends the ABP subunits contain repeated stretches of overlapping structures; these give the subunits more rigidity and allow them to hold actin filaments far apart.

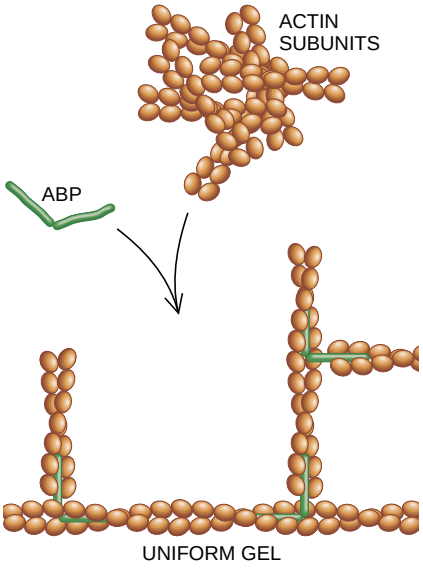
Using instruments that can measure the mechanical properties of gels, Paul A. Janmey of Harvard has shown that actin gels cross-linked by ABP are very strong and elastic: at the concentrations found in cells, actin and ABP could easily provide the stiffness of extended lamellae. Moreover, while working in our laboratory, Tadanao Ito of Kyoto University demonstrated that actin gels cross-linked by low concentrations of ABP can also retard water flow. Another scientist on sabbatical in our laboratory, Olle I. Stendahl of Linköping University Medical School in Sweden, used fluorescent antibodies to show that ABP molecules are localized in the cortical region of white blood cells. These findings all supported the theory that ABP helped to assemble actin filaments into gels in the cellular cortex.

Curious about the microscopic structure of actin gels cross-linked by ABP, Hartwig decided to examine them at high resolutions using a technique pioneered by John Heuser, now at Washington University. The method involves rapidly freezing specimens in liquid helium to preserve their structures. The ice in the specimens is removed by sublimation in a vacuum; the remaining material is shadowed with metals to make it visible under an electron microscope.

Hartwig showed that in the test tube, actin in the presence of ABP formed a uniform orthogonal network of randomly polarized filaments. On average, the filaments were one micron long and branched at perpendicular angles every 100 nanometers. He found a nearly identical network inside the lamellae of white blood cells; the ABP molecules were situated at the branch points between the filaments. This architecture contrasts with that of the filopodia, in which the actin filaments are all parallel, their barbed ends pointing away from the cell body. Thus, our studies strongly suggested that ABP was an excellent candidate to organize the elastic, spongelike actin gel of the leading lamella.

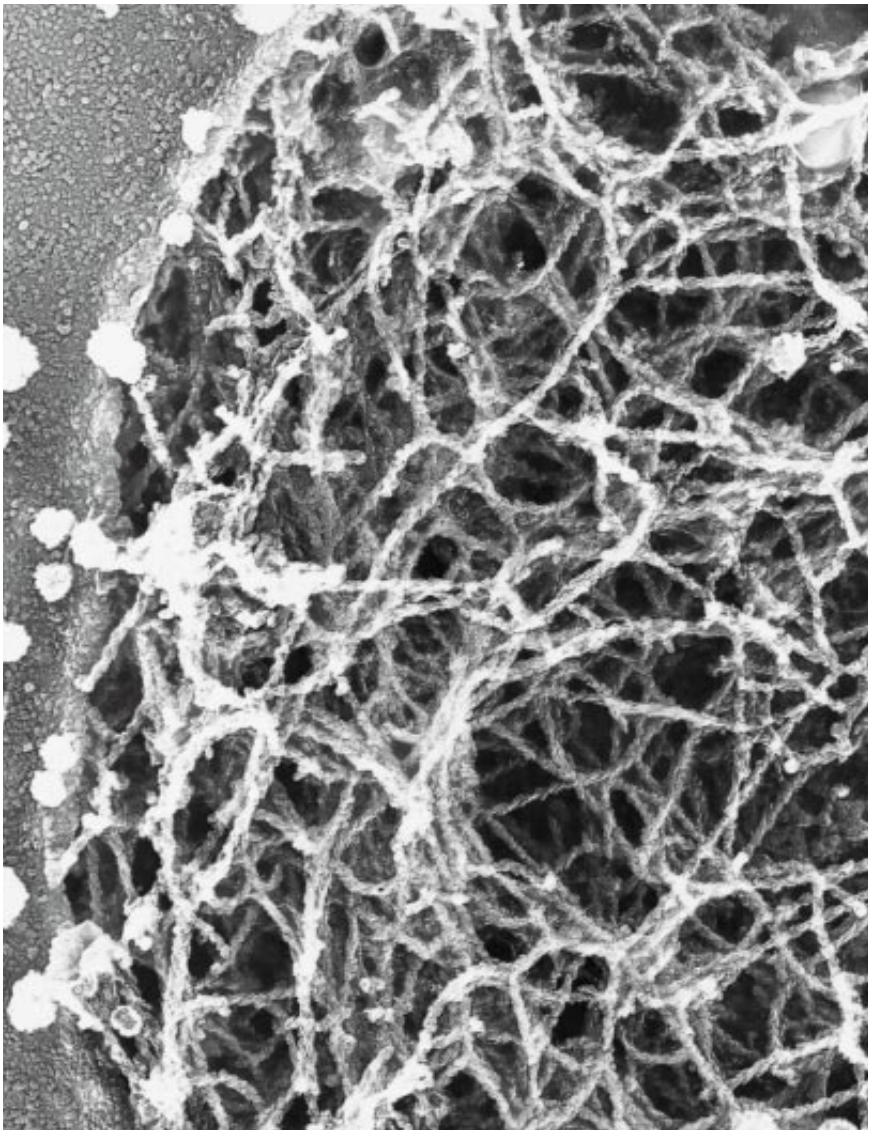
The best evidence for ABP's role came from asking, "What would happen if a cell did not have ABP?" In a study led by C. Casey Cunningham, our laboratory analyzed the protein composition of cell lines derived from the tumors of six patients with malignant melanoma. Three of those cell lines contained ABP, and three did not. The ABP-containing cells had the typical features of crawling cells: they extended glassy lamellae and moved toward chemoattractants. In contrast, the cells lacking ABP had normal filopodia but behaved as though their cortices were unstable. They did not put out a leading lamella and crawl in response to stimuli. Rather the cells sat in a state of dissonance as scores of unstable spherical projections, or blebs, extended and retracted over their entire surface. Normal cells produce blebs on occasion, but the ABP-deficient cells blebbed constantly.

Our interpretation of these observations was that in the ABP-deficient cells, the cortical gel is weak. When the sol flows in response to cellular contractions, its movement is poorly modulat-



ed and results in spherical protrusions. Inside a bleb, the actin filaments do not configure as a uniform gel, but they do eventually form a mass with sufficient coherence to be pulled back into the cell body. When we inserted a working gene for the ABP subunits into these defective cells, their permanent blebbing disappeared or diminished, and they developed the ability to crawl.

For a cell to crawl, the actin gel must remodel itself. In a cell at rest, about half of the total actin is not polymerized but is instead in the form of individual protein subunits that flow with the sol. Only at certain intracellular sites, in response to the right stimuli, does the actin polymerize. In a crawling cell the total amount of polymerized actin may remain constant: polymerization in one part is balanced by depolymerization elsewhere.



CROSS-LINKING of actin filaments into a gel is caused by ABP (actin-binding protein). Molecules of ABP organize random clusters of actin filaments into a highly uniform, nearly orthogonal, three-dimensional array (*top*). The micrograph (*bottom*) shows this spongelike network inside the leading lamella of a white blood cell.

As Oosawa established, the spontaneous polymerization of actin into filaments requires that two or three actin subunits first aggregate into a cluster, or nucleus. That event occurs infrequently, but once a nucleus has formed, it rapidly elongates through the addition of more subunits onto the ends of the incipient filament. A number of researchers have shown that the barbed end of the actin filament elongates much faster than does the pointed end.

Cells regulate actin assembly and disassembly with two general classes of control proteins. One of these types, of which three subclasses are known, binds predominantly or solely to actin subunits. Vivianne T. Nachmias and Daniel Safer of the University of Pennsylvania, working with the protein thymosin, determined that these proteins inhibit the spontaneous nucleation of actin subunits. The proteins also inhibit the addition of subunits to the pointed ends.

These actin subunit–binding proteins slow but do not prevent the addition of actin subunits to the barbed ends. By themselves, therefore, they cannot explain why so much actin remains unpolymerized in cells. The full explanation involves a second form of control exerted at the barbed ends by the other class of control protein.

Helen Lu Yin, now at the University of Texas at Dallas, and I discovered the first of these actin-binding proteins in extracts from white blood cells in 1979. In the presence of calcium concentrations that would be found inside cells stimulated to crawl, the protein "caps," or blocks, the barbed ends of actin filaments, thereby preventing the addition of more subunits. It also breaks the bonds holding together subunits in the filament—severing the filament and firmly affixing itself to the newly exposed barbed end. Because this protein drastically shortens the lengths of the actin filaments, it can convert an actin gel into a sol. We therefore named this protein "gelsolin."

Following the discovery of gelsolin, investigators in many laboratories recognized a variety of other actin-binding proteins that sever actin filaments or bind to the barbed ends of actin filaments, or both. These proteins fall into three subclasses. One is an extended family of capping proteins with primary structures related to that of gelsolin. Some but not all of these proteins also sever actin filaments. The structurally distinct second subclass is generally designated as Cap Z proteins; it was first identified independently in amoebas by Thomas D. Pollard and in blood plate-

lets by Shin Lin, both at Johns Hopkins University. The third subclass, initially discovered in brain tissue by James R. Bamburg of Colorado State University and Alan G. Weeds of the MRC, is abundant and weakly breaks actin filaments. Proteins of this subclass are named ADF, cofilin, depactin and actophorin.

Calcium causes proteins of the gelsolin family to stick to the barbed end of an actin filament, but as Joseph Bryan of Baylor College of Medicine has shown, simply removing calcium does not cause the gelsolin to release its grip. For a time, no one knew how this binding was reversed. Then in 1987 Janmey and I followed up on an observation by Ingrid Lassing and Uno Lindberg of Stockholm University concerning polyphosphoinositides. This class of phospholipid molecules is a common constituent of cell membranes and has been implicated in signal transduction within cells. Lassing and Lindberg had noticed that polyphosphoinositides could decrease the affinity of profilin, an actin subunit–binding protein that Lindberg had discovered in 1977, for actin subunits.

We demonstrated that these phospholipids had a twofold effect on gelsolin: they specifically inhibited its filament-severing activity and also caused it to dissociate from the barbed ends of actin filaments. A variety of experiments conducted in many laboratories around the world have further shown that polyphosphoinositides inhibit the actin-binding activity of nearly all proteins that cap and sever actin filaments.

All this information points to a model for regulated cell crawling—one that integrates both stimulus-induced signal transduction and the remodeling of the actin gel in the cellular cortex. When chemoattractants and other agents stimulate cells, enzymes in the cell membrane begin synthesizing or breaking down polyphosphoinositides. One consequence of the breakdown reaction is the release of calcium into the sol filling the cell from membrane-bound storage vesicles. Because calcium activates the actin-capping proteins of the gelsolin family, polyphosphoinositide degradation would lead to actin disaggregation.

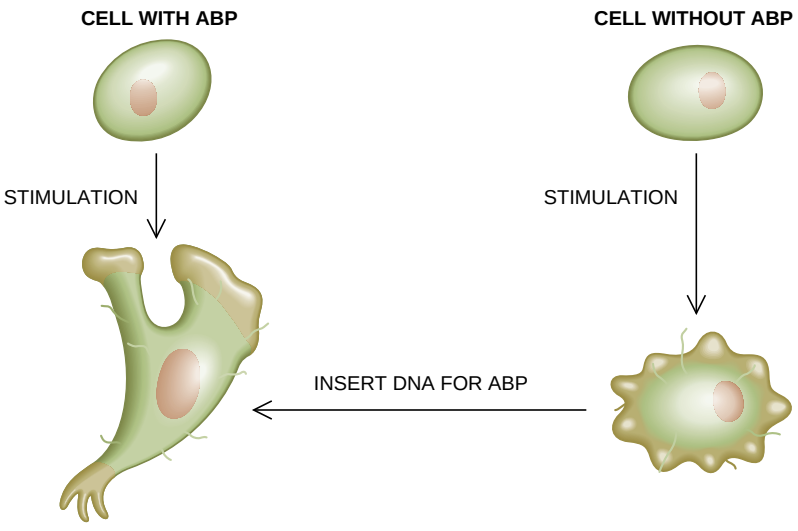
On the other hand, the synthesis of polyphosphoinositides would cause the uncapping of actin filaments near the plasma membrane. It would thereby promote the assembly of actin into elongating filaments. The effectiveness of polyphosphoinositides for uncapping actin filaments is influenced by their chemical environment. Cold may cause the actin in platelets to gel irreversibly because it induces phase changes in the cell membrane and permanently alters the presentation of these phospholipids.

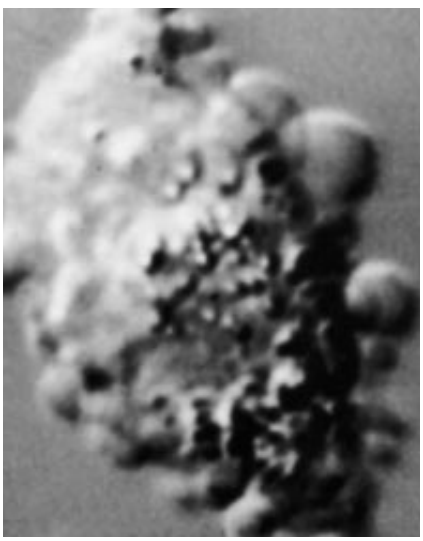
For cells to crawl, it is not enough that actin form a gel: it must also be possible for myosin to pull on that gel. As I have mentioned, calcium activates the contraction of cortical actin by phosphorylating myosin; it also partially dissolves the actin gel by activating gelsolin and related proteins. The gel must disaggregate sufficiently to allow myosin to move the actin filaments, yet not so extensively that the gel becomes entirely liquid.

D. Lansing Taylor, now at Carnegie Mellon University, has termed this coordinated event "solation-contraction coupling." He and his colleagues have used mixtures of actin filaments, ABP and gelsolin to demonstrate the feasibility of this mechanism. Actin subunits and capped short filaments released from the solating gel percolate through the lamella to the protruded membrane. There polyphosphoinositides uncap the filaments, which elongate by addition of subunits and incorporate into the gel.

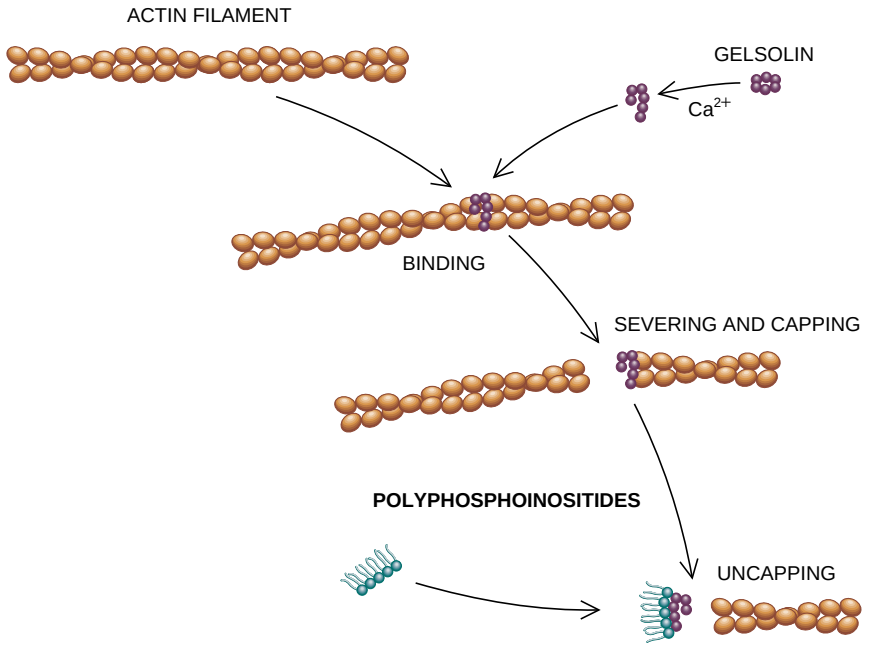
Observations in living cells support this picture of regulated actin assembly and its relation to cell crawling. When Yu-li Wang of the Worcester Foundation for Experimental Biology microinjected fluorescently labeled actin into crawling fibroblasts, he saw that the labeled actin incorporated itself into filaments at the cell's leading edge. John Condeelis of the Albert Einstein College of Medicine in Bronx, N.Y., Sally H. Zigmond of the University of Pennsylvania and Hartwig of Harvard have found that stimulating cells with chemoattractants leads to the uncapping of the barbed ends of actin filaments. Hartwig has documented that the fragmentation of actin filaments in the margin of platelets depends on a rise in intracellular calcium. These findings strongly implicate severing proteins of the gelsolin family in crawling.

The mechanisms involving calcium, phospholipids and actin-binding proteins that I have discussed are clearly not the only ones at work in the regulation of intracellular actin assembly. For instance, actin subunits bind ATP or adenosine diphosphate (ADP). ATP-containing subunits polymerize more efficiently than those that contain ADP. During polymerization, bound molecules of ATP are also enzymatically converted to ADP, liberating energy. When the actin subunits dissociate from filaments,





ABP IS ESSENTIAL for human cell crawling. Normal cells, when stimulated to crawl, extend a leading lamella. Abnormal cells without ABP, however, instead randomly produce many blebs, or small spherical protrusions. (A blebbing cell is shown at right.) If the abnormal cells are treated with DNA that enables them to make ABP, they assume a normal crawling shape.



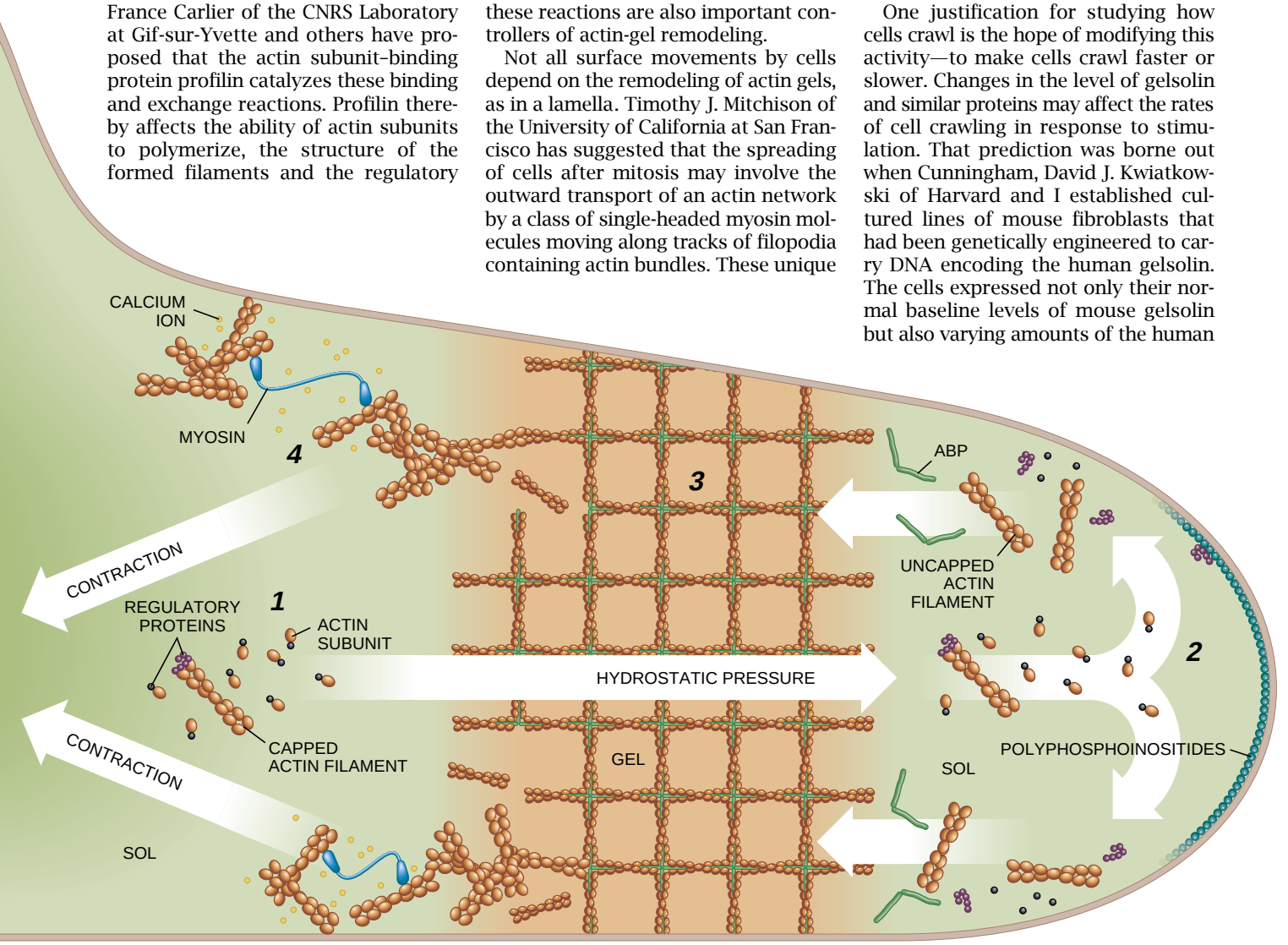
GELSOLIN is one of the proteins that regulates the assembly and disassembly of actin filaments. In the presence of calcium, gelsolin severs actin filaments and "caps" their barbed end, preventing the addition of new subunits there. Lipid molecules called polyphosphoinositides, which are found in the cell membrane, can remove gelsolin from actin filaments and allow them to elongate.

they swap their ADPs for ATPs. Marie-France Carlier of the CNRS Laboratory at Gif-sur-Yvette and others have proposed that the actin subunit–binding protein profilin catalyzes these binding and exchange reactions. Profilin thereby affects the ability of actin subunits to polymerize, the structure of the formed filaments and the regulatory influence of other proteins. As such, these reactions are also important controllers of actin-gel remodeling.

Not all surface movements by cells depend on the remodeling of actin gels, as in a lamella. Timothy J. Mitchison of the University of California at San Francisco has suggested that the spreading of cells after mitosis may involve the outward transport of an actin network by a class of single-headed myosin molecules moving along tracks of filopodia containing actin bundles. These unique myosins were discovered by Pollard and Edward D. Korn of the NIH.

The mechanism of filopodial extension appears to differ from that of lamellar protrusion. Lewis G. Tilney of the University of Pennsylvania first showed in the early 1970s that filopodial extension involves actin assembly. George F. Oster of the University of California at Berkeley has presented a mechanism for these protrusions that he calls the Brownian ratchet. In his model, thermal fluctuations in the cell membranes are harnessed to direct the assembly of actin and force the protrusion of a filopod.

One justification for studying how cells crawl is the hope of modifying this activity—to make cells crawl faster or slower. Changes in the level of gelsolin and similar proteins may affect the rates of cell crawling in response to stimulation. That prediction was borne out when Cunningham, David J. Kwiatkowski of Harvard and I established cultured lines of mouse fibroblasts that had been genetically engineered to carry DNA encoding the human gelsolin. The cells expressed not only their normal baseline levels of mouse gelsolin but also varying amounts of the human



protein. Tests revealed that the locomotion of these cells increased in proportion to their concentration of gelsolin.

These results proved that tinkering with the intracellular machinery of cell crawling can affect rates of locomotion, at least in the laboratory. Useful applications of more refined manipulations are not hard to imagine. Speeding up the movements of fibroblasts, for example, might accelerate the healing of wounds. Conversely, if we could partially inhibit cell crawling, we might impede the ravages of destructive inflammation by white blood cells or of coronary artery thrombosis mediated by activated platelets.

A more immediate practical spin-off of this research program may have emerged in 1979. Working independently, Astrid Fagraeus and René Norberg of Uppsala University and Christine Chaponnier and Giulio Gabbiani of the University of Geneva discovered that substances in blood plasma cause filamentous actin to depolymerize. The molecular basis of this activity was subsequently shown by others to reside in two plasma proteins that work cooperatively: Gc globulin, a genetically polymorphic protein that binds to actin, and a secreted form of gelsolin.

As several investigators have documented, injured animals and humans often have extracellular actin in their blood; in addition, their levels of Gc globulin and plasma gelsolin are depleted. Stuart E. Lind of Harvard and John G. Haddad of the University of Pennsylvania have also found that because of its complex effects on blood coagulation, extracellular actin can be toxic to tissues—even lethal. Gc globulin and plasma gelsolin may therefore be parts of an actin-scavenging system.

Recently we have seen that such actin scavenging may be important in cystic fibrosis, the most common hereditary disorder affecting people of European descent. For poorly understood reasons, the underlying genetic defect, a mutation in the gene for a chloride-transport regulator protein, leads to abnormal secretions in the airways of the lungs. As a result, the lungs become inflamed and infected with bacteria. In a process that involves prodigous crawling by neutrophils, the airways then fill with white blood cells that subsequently degenerate. This purulent matter causes the lung mucus to gel into a highly viscous mass that can gradually suffocate the cystic fibrosis patient.

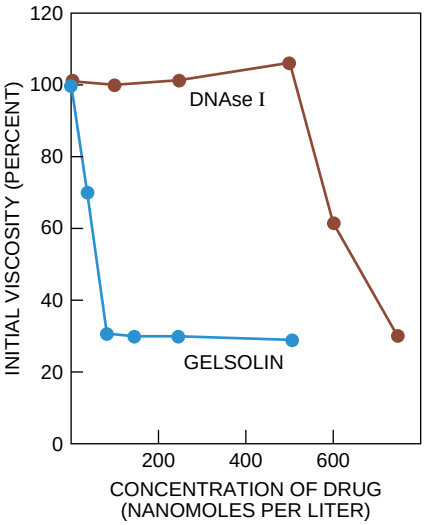
CRAWLING by the proposed mechanism depends on the regulated assembly and disassembly of actin filaments. The sol in a cell body contains actin subunits bound to regulatory proteins that prevent them from assembling (*1*). When the cell is stimulated, hydrostatic force carries these subunits through the weakened gel and into the protruding lamella. Lipids in the membrane free the subunits from the regulatory proteins (*2*). Actin filaments then rapidly begin to form and, with the assistance of ABP, form a gel (*3*). At the trailing edge of the gel, calcium ions reactivate the actin-severing proteins, which loosen the actin network enough for myosin molecules to pull on it (*4*). Subunits from the disassembling gel are reused as new protrusions form. In this way, the cell pulls itself forward continuously.

CLEAVING ACTIN might be a key to thinning the heavy mucus that clogs the airways of cystic fibrosis patients. The viscosity of this mucus has generally been attributed to large amounts of polymerizing DNA from dying cells, but actin filaments could also account for much of it. In tests, the actin-severing protein gelsolin was able to thin the mucus of patients more efficiently at low concentrations than a type of DNA-cleaving enzyme did.

The gelation of the mucus in cystic fibrosis patients has long been ascribed to polymers of DNA released from the nuclei of the dying neutrophils. To treat the disease, medical investigators have therefore developed a genetically engineered form of an enzyme, deoxyribonuclease I (DNAse I), which was recently approved for clinical use. This DNAse I diminishes the consistency of patients' airway secretions in vitro. According to reports, when this drug is inhaled, it can improve pulmonary function. Ostensibly, the drug works by enzymatically cleaving long DNA polymers.

Nevertheless, our research on the gelation of actin suggests that DNAse I may achieve its beneficial ends by a different mechanism. In 1963 Lindberg purified a protein that inhibited the natural form of DNAse I. Ten years later, while on sabbatical at Cold Spring Harbor Laboratory, he and Elias Lazarides determined that this inhibitory protein was actin. Like Gc globulin, DNAse I tightly binds to actin subunits. If DNAse I is present in sufficient quantities, it can depolymerize actin filaments by blocking the addition of subunits to shrinking filaments. My colleagues and I therefore decided to consider whether filamentous actin—which may be as abundant in white blood cells as DNA is—might contribute significantly to the gelation of mucus in cystic fibrosis.

Our studies found that actin is present in patients' sputum, where it would presumably inhibit the DNA-cleaving activity of DNAse I. We also showed that the addition of plasma gelsolin reduced the viscosity of the mucus. Indeed, gelsolin appears to be even more efficient at dissolving the mucus than DNAse I is (which we had anticipated, given that gelsolin severs actin filaments).



Because DNAse I and gelsolin work by different mechanisms, it is possible that together their therapeutic effects might be synergistic. Gelsolin is a normal extracellular constituent of the body. Its administration to the airways should theoretically be nontoxic and should not evoke an immune response. Perhaps clinical investigations can someday turn these findings into an appropriate treatment for cystic fibrosis. If so, this work will add to biologists' previous evidence that basic research on such arcane matters as the gelation of actin in crawling cells can lead to medical advances.

FURTHER READING

THE EXTRACELLULAR ACTIN-SCAVENGER SYSTEM AND ACTIN TOXICITY. W. M. Lee and R. M. Galbraith in *New England Journal of Medicine,* Vol. 326, No. 20, pages 1335–1341; May 14, 1992.

LIFE AT THE LEADING EDGE: THE FORMATION OF CELL PROTRUSIONS. J. Condeelis in *Annual Review of Cell Biology,* Vol. 9, pages 411–444; 1993.

ON THE CRAWLING OF ANIMAL CELLS. Thomas P. Stossel in *Science,* Vol. 260, pages 1086–1094; May 21, 1993.

PHOSPHOINOSITIDES AND CALCIUM AS REGULATORS OF CELLULAR ACTIN ASSEMBLY AND DISASSEMBLY. Paul A. Janmey in *Annual Review of Physiology,* Vol. 56, pages 169–191; 1994.

REDUCTION IN VISCOSITY OF CYSTIC FIBROSIS SPUTUM IN VITRO BY GELSOLIN. C. A. Vasconcellos et al. in *Science,* Vol. 263, pages 969–971; February 18, 1994.

Solving the Paradox of Deep Earthquakes

For decades, geophysicists have known that earthquakes should not occur at depth inside the earth. But they do. Finally, we know how and why these events happen

by Harry W. Green II

On June 8 of this year, a great earthquake rumbled through the earth's mantle more than 600 kilometers below Bolivia. It was the largest earthquake ever recorded at such depths and the biggest of any kind in the past 15 years. The tremors were felt as far away as Toronto. No temblor in history had shaken the earth so far from its epicenter.

The event was truly spectacular and yet paradoxical as well. Although deep earthquakes are as regular as clockwork, they should not, in theory, be possible. The very existence of deep earthquakes has teased geophysicists since their discovery in 1927. Five years ago my colleagues and I in my laboratory at the University of California, first at Davis and now at Riverside, began to unravel the solution to this puzzle. This article gives an account of that discovery and of the new theory of earthquakes that has flowed from it.

Most earthquakes occur within a few tens of kilometers of the earth's surface by the familiar processes of brittle fracture and frictional sliding—the same mechanisms by which glass breaks and tires squeal on pavement. Yet almost 30 percent of all earthquakes occur at depths exceeding 70 kilometers, where the pressure reaches upward of two gigapascals (20,000 times that of the atmosphere at sea level); nearly 8 percent happen at depths greater than 300 kilometers, where the pressure is greater than 10 gigapascals. At such high pressures, rock will flow at lower stresses than those at which it will break or slide along a preexisting fault. Earthquakes at depth, then, would seem impossible.

Nevertheless, deep earthquakes do occur, exclusively in thin, planar zones in the earth that begin underneath oceanic trenches and angle down into the mantle. The theory of plate tectonics posits that these locations mark subduction zones, where the cold uppermost layer of the earth (the lithosphere, 50 to 100 kilometers thick) sinks into the mantle. In doing so, it provides the return flow that compensates for the upwelling of molten material and creation of lithosphere at ocean ridges. In these zones, earthquakes show an exponential decrease in frequency from the surface to about 300 kilometers deep. Then their frequency increases again, peaking at 550 to 600 kilometers deep. Finally, earthquakes cease entirely at approximately 680 kilometers deep.

Because the frequency of earthquakes steadily declines down to about 300 kilometers, most geophysicists believe that events originating between 70 and 300 kilometers below the surface (termed intermediate-focus earthquakes) are produced by a mechanism simply related to brittle fracture and frictional sliding. Deep-focus earthquakes (below 300 kilometers), however, follow an entirely different pattern and therefore must stem from a separate mechanism. For more than six decades, the details of this mechanism remained elusive.

Years of study did provide intriguing information about subduction zones. Near the earth's surface, rocks contain minerals that exhibit a relatively loose packing of atoms. As the pressure on them increases at greater depths within the mantle, the atoms reorganize and yield minerals having progressively greater density. The first such transformation occurs in most parts of the mantle at a depth of about 400 kilometers. In the reaction, olivine, the most abundant mineral of the upper mantle, becomes unstable and changes into a phase having a spinel (cubic) structure that is 6 percent more dense than the original mineral. This shift causes an abrupt increase in seismic velocity at this depth. At 660 kilometers, the spinel form itself becomes unstable and decomposes into two phases, which together are an additional 8 percent more dense. The reaction induces another sharp rise in seismic velocity, marking the boundary between the upper and lower mantles.

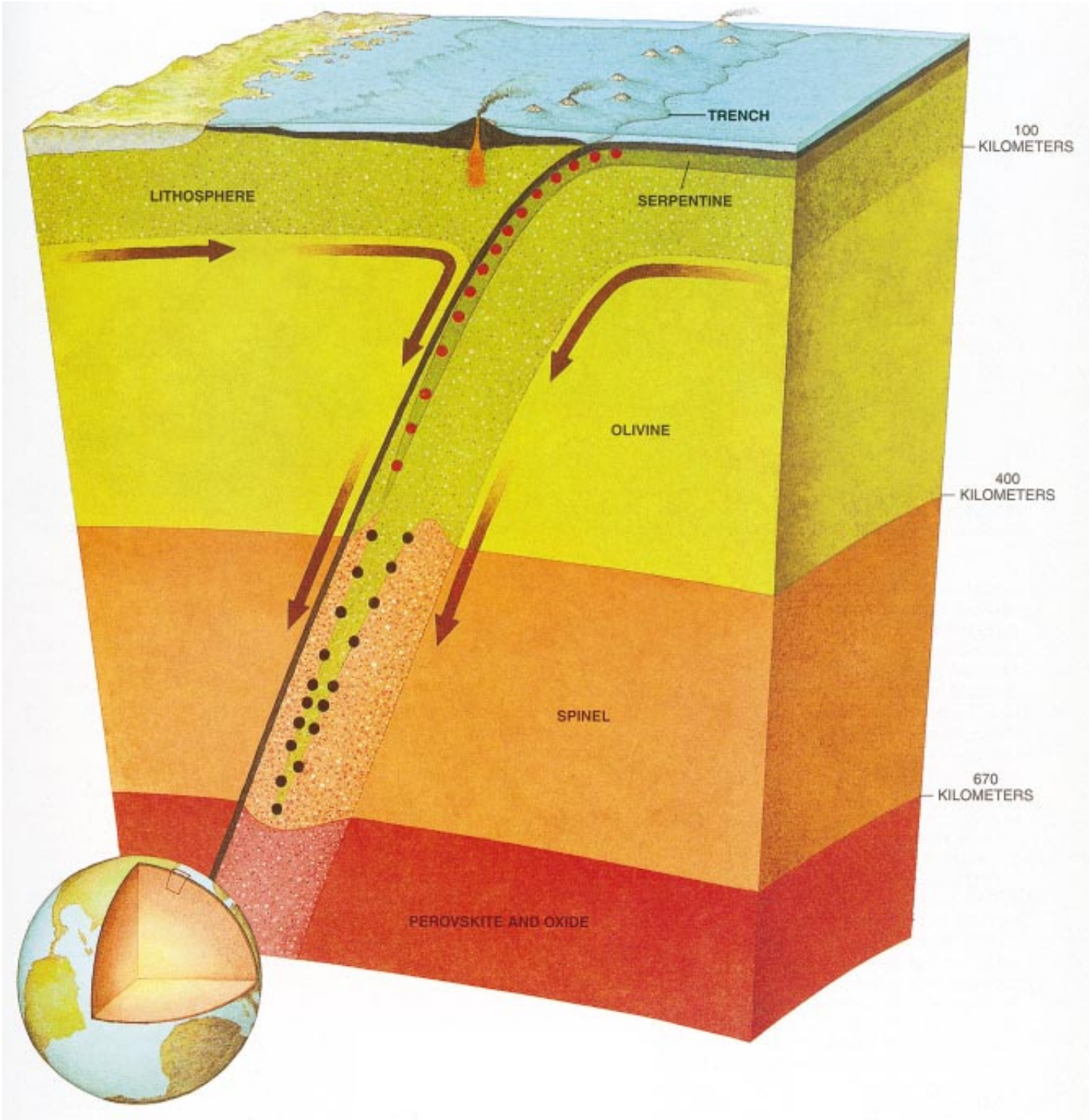
The temperature is lower in a subducting slab. Under these conditions, the spinel structure becomes stable at somewhat lower pressures than normal and remains so until reaching slightly higher pressures than normal. Hence, the spinel stability field extends from a depth of about 300 kilometers to a depth of about 700 kilometers. This is exactly the region in which deep-focus earthquakes occur.

Because of this correlation, one of the recurring explanations over the years has been that the distribution of deep-focus earthquakes relates in some unknown way to these phase transformations. Most early suggestions centered around the fact that the reactions involve densification. Several researchers proposed that a sudden transforma-

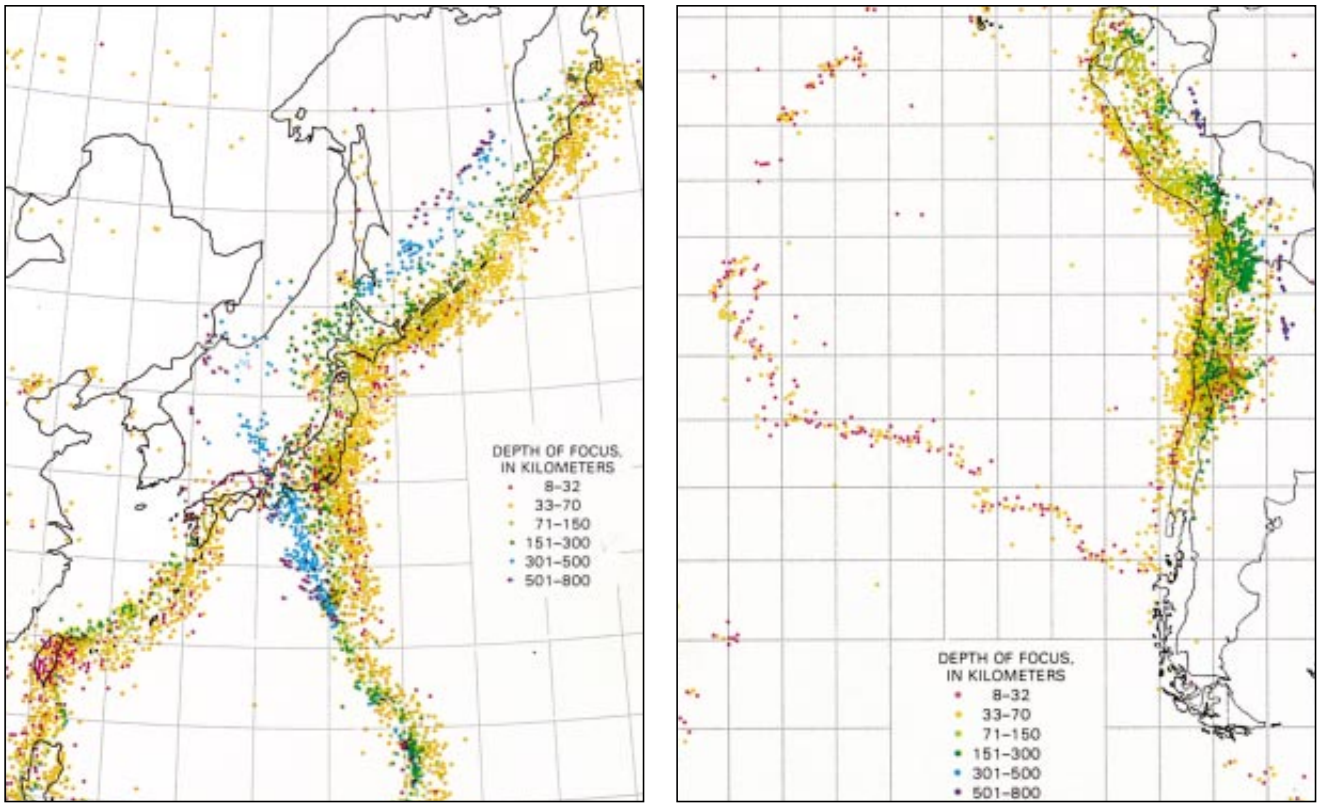
HARRY W. GREEN II is a professor at the University of California, Riverside, where he directs the Institute of Geophysics and Planetary Physics. In 1968 he received a Ph.D. in geology and geophysics from the University of California, Los Angeles. Currently he investigates the rheology of geologic materials and how stress affects polymorphic phase transformations. He is a fellow of the Mineralogical Society of America and a member of many other professional organizations. He has published more than 70 papers and offers counsel to the National Science Foundation, the Department of Energy and other such agencies.

tion of a significant volume of olivine to spinel would produce an implosion that could radiate the required seismic energy. Later studies refuted this hypothesis, however, showing that the geometric pattern of seismic energy that radiates from deep earthquakes is indistinguishable from that of shallow ones. Moreover, it indicates that movement takes place along a fault.

So what does cause deep earthquakes, and why do the events correlate with the spinel stability field? Direct experimentation of any kind at the extraordinary pressures of the earth's deep interior has become possible only in the past three decades. In 1976 Chien-Min Sung and Roger G. Burns of the Massachusetts



SUBDUCTION ZONES, where tectonic plates meet, are the only places where intermediate- and deep-focus earthquakes occur. Shallow earthquakes happen throughout the world when the brittle rock in the lithosphere fractures and slides. In a cold subducted slab, different mechanisms promote deeper events. Intermediate-focus temblors (*red dots*) occur when serpentine (olivine and water) is dehydrated as it descends into the mantle. Deep-focus quakes (*black dots*) result from the growth and spread of dense microstructures around the margins of the metastable olivine wedge that extends below 400 kilometers. At 700 kilometers, any remaining olivine decomposes silently, and all earthquake activity stops.



SEISMICITY mapped near Japan (*left*) and South America (*right*) from 1971 to 1986 shows that most earthquakes are shallow, originating at depths less than 70 kilometers below the earth's surface. Intermediate-focus earthquakes, those below 70 kilometers, and deep-focus events, those below 300 kilometers, together make up only a third of all earthquakes.

Institute of Technology demonstrated that for temperatures and pressures expected in the cold core of a subduction zone, the transformation of olivine to spinel would probably be kinetically inhibited, even on a time scale of tens of millions of years (more recently, David C. Rubie and colleagues at the Bayerisches Geoinstitut in Germany have confirmed these results and established that metastable olivine should persist in rapidly subducting lithosphere).

In the same year that Sung and Burns published their initial results, J. Rimas Vaisnys and Carol C. Pilbeam of Yale University suggested that a faulting instability might be possible during the transformation from olivine to spinel under certain conditions. In particular, they appealed to a thermal runaway (an exothermic reaction releases heat, which speeds the reaction rate even more, and so on) and a marked decrease in crystal size, important characteristics that I will discuss further.

Also in the late 1970s and early 1980s a controversy arose concerning the exact mechanism by which olivine transforms to spinel. In addition to the silicate olivine of the earth's mantle, $(Mg, Fe)_2SiO_4$, the olivine-spinel transformation takes place in several chemical systems, including germanate olivine, Mg_2GeO_4. Because the germanium atom in this compound is larger than a silicon atom, the transformation happens at much lower pressures than it does in silicate olivine. Work in my laboratory using the germanate system agreed with the earlier observations of Sung and Burns—namely that the transformation occurred by the nucleation and growth of spinel crystals on olivine grain boundaries. Studies elsewhere, though, supported a different kind of mechanism, in which the crystal lattice sheared. The differences between the various experiments caused me to propose in 1984 that both mechanisms must exist and that stress probably determined which one would operate under a given set of conditions.

It was important to resolve the issue because understanding the various aspects of mantle dynamics (including deep earthquakes) depends on knowing the mechanism responsible for this transformation. Thus, in 1985, Pamela C. Burnley (who was then a graduate student beginning her Ph.D. research) and I began investigating the effect of stress on the transformation. It was not then (and still is not) possible to perform deformation experiments and measure stress at the very high pressures under which this transformation takes place in the silicate system. Therefore, Burnley (who is now at the University of Colorado) and I continued to use magnesium germanate samples, because the pressure needed to induce the transformation is readily accessible in my experimental deformation machinery.

We prepared and deformed small samples of a synthetic "rock" of this composition within the stability field of the spinel polymorph. The work confirmed that stress levels determine which of the two mechanisms will run. At low temperatures, under conditions too cold for the reaction to run by nucleation and growth of new crystals, our specimens were very strong. They transformed only when high stress caused the crystal lattice to shear into thin lamellae of the denser phase. At high temperatures, however, the nucleation and growth mechanism ran quickly, and so the specimens were much weaker. In this case, the high stress that produced the shearing mechanism was never reached.

These results resolved the controversy over how olivine transforms into spinel. But the stresses required to produce the shearing mechanism are so high that only the nucleation and

growth mechanism should operate in the earth. Moreover, we found no faulting instability associated with the shearing mechanism. Thus, it could be ruled out as a possible mechanism for deep earthquakes as well.

At the same time that Burnley was conducting these experiments, Stephen H. Kirby of the U.S. Geological Survey in Menlo Park, Calif., reported some anomalous results. He was performing faulting studies of two minerals conducted near or above the pressure at which a densification reaction might be expected. Although he found no direct evidence of such a reaction, Kirby proposed that incipient transformation to the stable phases might have caused the faulting he observed. Like Vaisnys and Pilbeam 10 years earlier, he suggested that a faulting instability might operate in the earth's mantle during the transformation from olivine to spinel.

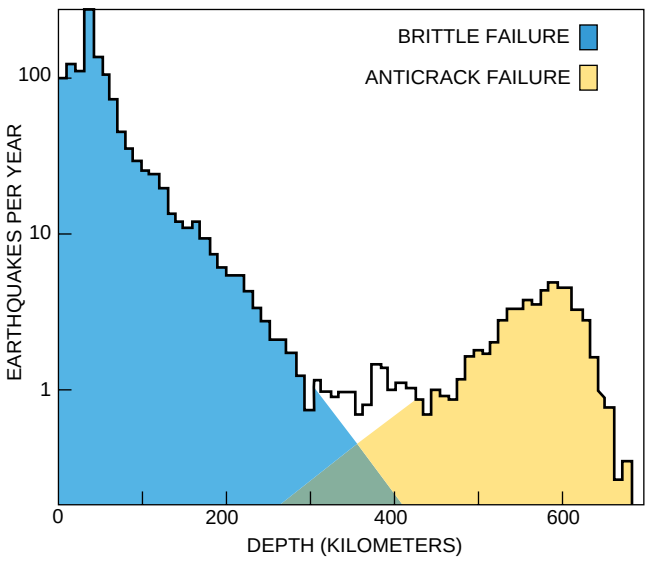
Although we had yet to witness this predicted instability, Burnley and I reasoned that if such an instability existed, it had to involve the nucleation and growth mechanism. Furthermore, the instability had to appear only in the narrow temperature interval between the two ranges tested during our earlier work. Consequently, we deformed specimens under conditions for which nucleation of the spinel phase is just possible on the time scale of the experiment. Bingo! These specimens exhibited an abrupt drop in the amount of stress they could support and developed one or more spinel-lined faults cutting through them.

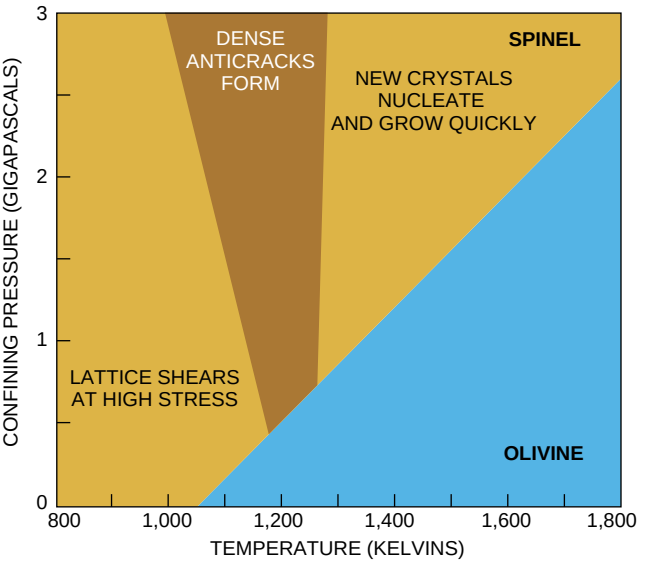
Detailed examinations revealed a unique set of microstructures within these faulted specimens. Early on, during experiments conducted within the narrow faulting "window," microscopic packets of the high-density phase formed and grew on the olivine grain boundaries. These packets exhibited three critical characteristics: they looked like filled cracks; they ran perpendicular to the stress field; and they contained extraordinarily small crystals of spinel (approximately 10^{-5} millimeter in diameter). The first two characteristics are tantalizingly similar to features that develop in brittle materials before they break. The third offered a potential answer as to how faults can form and slide at high pressures.

From these three characteristics, we formulated a theory of transformation-induced faulting that is analogous to brittle shear fracture but that differs fundamentally in its microphysics. In brittle shear fracture, as the stress rises, large numbers of microscopic tensile cracks open parallel to the maximum compressive stress (S_1). These features are referred to as Mode I cracks because the displacements across them are perpendicular to the plane of the crack. As loading continues, the number and density of Mode I microcracks increase rapidly until the material begins to lose its strength locally. At that time, the microcracks cooperatively organize to initiate shear fracture, and the specimen fails in a fraction of a second. A "process zone" of tensile (Mode I) microcracks develops in front of the growing fault and leads it through the material. The important point here is that the fault is not a primary failure process; it must be prepared for and led by Mode I microcracks. Because pressure severely inhibits the expansion that takes place when tensile microcracks open, brittle failure cannot occur at depth in the earth.

In our high-pressure faulting experiments, we observed the growth of microscopic lenses of spinel in place of microcracks. The lenses are shaped very much like open tensile cracks, but they have the opposite orientation—they form perpendicular to S_1. The spinel phase is more dense than olivine; hence, the displacements of the lens boundaries move inward toward the plane of the lens. Therefore, the lenses are Mode I features like tensile cracks. Because the displacement of their boundaries is reversed, however, concentrations of compressive stresses, rather than tensile stresses, develop at their tips. It is the tensile stresses at the tips of opening cracks in brittle materials that cause them to orient themselves parallel to S_1; similarly, the compressive stresses at the tips of the lenses in our specimens cause them to orient themselves perpendicular to S_1.

Thus, in every way these features are the inverse of cracks—in a word, they are anticracks, a concept advanced in 1981 in a different context by Raymond Fletcher of Texas A&M University and David D. Pollard of Stanford University. Because of the remarkable similarities between the two Mode I features, we concluded that the microanticracks that precede failure in our experimental spec-

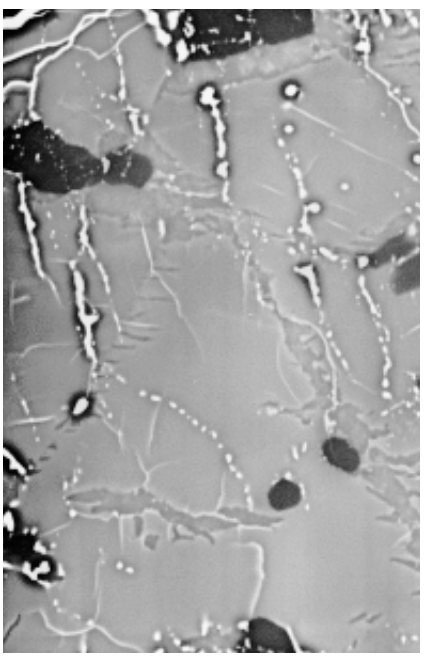




FREQUENCY of earthquakes corresponds closely to the depths at which olivine undergoes phase transformations (*left*). A minimum number of events occur at roughly 400 kilometers, where olivine transforms into a denser spinel (or cubic) phase. No earthquakes arise below a depth of 700 kilometers, where spinel decomposes. Pressure and temperature govern these reactions (*right*). In a germanate olivine at low pressures and high temperatures, olivine is stable, whereas at high pressures or low temperatures, the denser spinel is stable. Anticrack faulting occurs only in a narrow temperature "window."

imens must play the same role in high-pressure faulting as do microcracks in brittle fracture.

The third critical characteristic of our faulted specimens, the very fine grained spinel in the anticracks, gave us insight as to how anticracks can provide a fundamental weakening step and why the process can occur at high pressure. Extremely fine grained materials exhibit a remarkable flow property called superplasticity. Such materials flow by sliding on the grain boundaries between the crystals. This flow is somewhat like the deformation of a bag of sand, but with the all-important difference that the grains of sand are rigid. Therefore, they must slide up and over one another. As gaps open up between sand grains, the dilation must work against the ambient pressure. Hence, this process, like brittle failure, is severely inhibited by pressure. In contrast, grain-boundary sliding is a plastic process in which crystal defects called grain-boundary dislocations move. No expansion happens (as in the granular flow of sand), and so pressure has little inhibitory effect. We postulated that the fine-grained spinel within the anticracks is much weaker than the host olivine and has this "superplastic" flow capacity.



MICROSCOPIC LENSES of dense spinel-phase olivine (*white*) weakened germanate olivine specimens. The rocks were deformed at temperatures at which spinel crystals nucleate sluggishly. At lower temperatures, high stress can induce a smooth transformation from olivine to spinel-phase olivine. Likewise, at higher temperatures, spinel crystals nucleate easily and quickly.

From these observations we formulated the following hypothesis. During loading, under conditions for which the spinel phase grows with difficulty, olivine transforms to spinel. The transformation takes place as new crystals form by repeated nucleation adjacent to one another where stress concentrates. In a nonhydrostatic stress field, the developing packets of spinel tend to grow perpendicular to S_1. This preference leads to their lens-shaped morphology and alignment. These Mode I microanticracks initially form scattered throughout the specimens. But because the fine-grained spinel aggregates within the microanticracks are much weaker than the large olivine crystals, once enough of them have formed, the specimen loses its strength locally.

At this critical stage, large stress concentrations develop around the region of incipient failure, and the growth of anticracks accelerates. Preexisting microanticracks then link up and empty their superplastic contents into the developing fault zone, providing a lubricant along which the fault can slide. The process continues ahead of the tip of the growing fault zone and thereby provides the superplastic material needed to lubricate the fault. The anticracks must grow very rapidly to produce this faulting. We postulated that the speed of their growth resulted from a thermal feedback mechanism: the nucleation of spinel in the anticracks releases heat that locally increases the temperature, which increases the nucleation rate, which raises the temperature further, prompting faster nucleation and leading to catastrophic failure.

Burnley and I published the essence of this model in *Nature* in October 1989, and much of the time in my laboratory since then has been spent testing various aspects of the theory. Happily, it has survived all our scrutiny thus far. In one very important test, we investigated whether energy is radiated elastically during anticrack faulting. Obviously, if anticrack faulting is "silent," it cannot be responsible for earthquakes because the shaking we experience is caused by the arrival of "noise" emitted during the failure process. Because our specimens were small and located deep within the deformation apparatus (which itself produces general background noise), we could not hear the sound emitted during the faulting process.

To overcome this difficulty, I established a collaboration with Christopher H. Scholz of the Lamont-Doherty Earth Observatory of Columbia University, who investigates brittle fracture in the earth. Scholz attaches sensitive piezoelectric transducers to his apparatus to "listen" to the acoustic emissions that precede and accompany brittle failure. We modified one of my high-pressure deformation apparatuses to reduce noise and, working with Tracy N. Tingle and Thomas E. Young from my lab and Theodore A. Kozynski from Scholz's lab, successfully detected acoustic emissions from samples of Mg_2GeO_4 during failure.

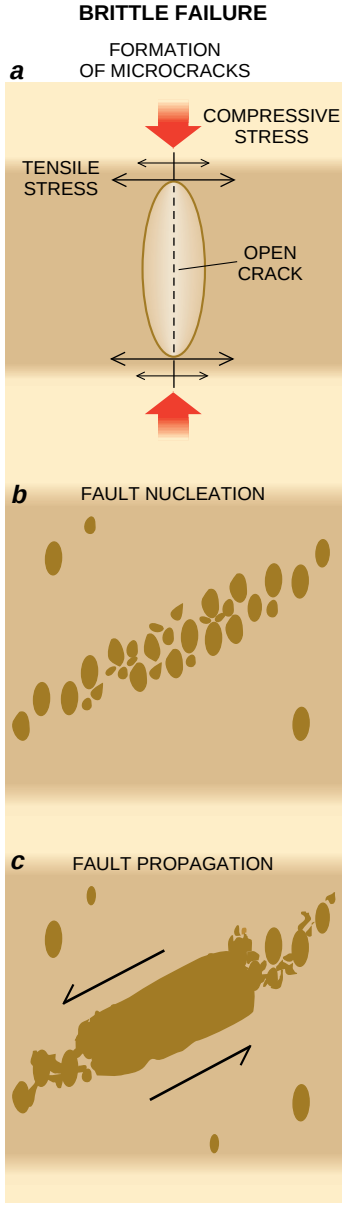
Tingle and I also investigated the flow strength of Mg_2GeO_4 spinel when the crystals are comparable in size to the olivine crystals of the starting material. We then compared that strength to the resistance against sliding present on anticrack-induced faults. Whereas this resistance is much less than the flow strength of the olivine specimens before failure, the flow strength of coarse-grained spinel is twice as great as that of olivine. As a result, one cannot explain the weakness of the fault zones in our specimens by simply replacing olivine with spinel; the flow mechanism must also change. The only known mechanism that can provide such weakening is superplastic flow, consistent with our original speculation.

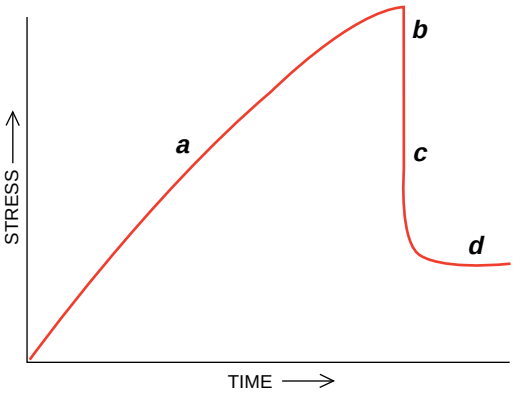
These tests established beyond doubt that anticrack faulting was a new failure mechanism distinct from brittle failure. Nevertheless, they had one major flaw. We conducted all these experiments on germanate olivine, not the silicate olivine found in the mantle. Of course, as mentioned earlier, none of this work could have been done on silicate olivine; it is still not possible to measure stresses at the high pressures needed to reach the spinel stability field in the silicate system.

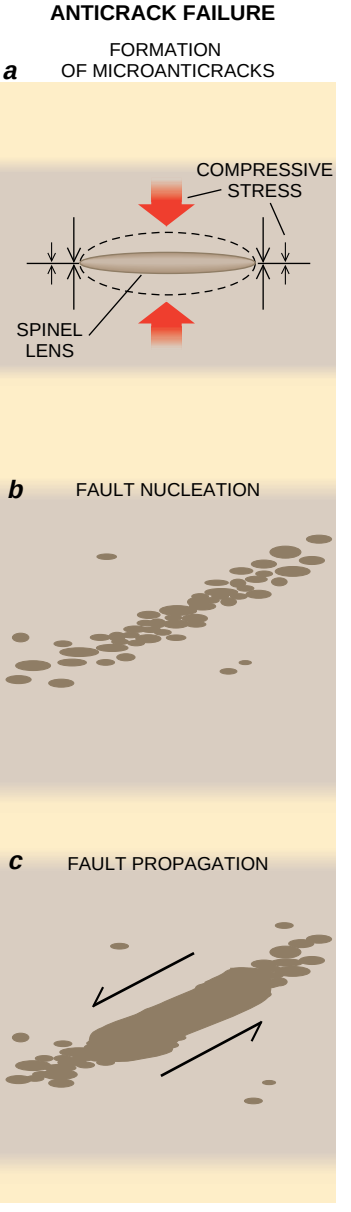
David Walker, also at Lamont, then suggested to Scholz and me that we attempt crude experiments on mantle olivine in his multianvil apparatus, a machine that can attain the requisite pressures to transform the silicate. Such a device had never before been used for deforming mineral specimens, but we decided to follow the advice. Our philosophy was that if anticrack faulting truly gives rise to deep-focus earthquakes, it must operate in real olivine. The microstructures we observed in the germanate specimens could guide us to uncover the conditions under which instability would develop in the silicate. The approach worked beyond our wildest dreams; after only four trials, we produced faulting and characteristic anticrack microstructures in mantle olivine at a pressure of 14 gigapascals.

Despite the attractive properties of the anticrack faulting mechanism, it can operate in the earth only if olivine

Brittle versus Anticrack Failure





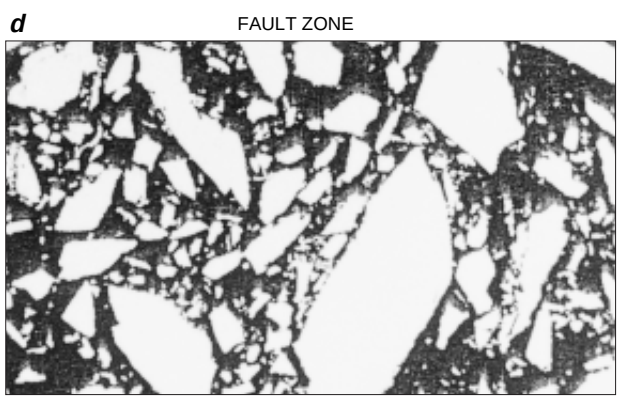


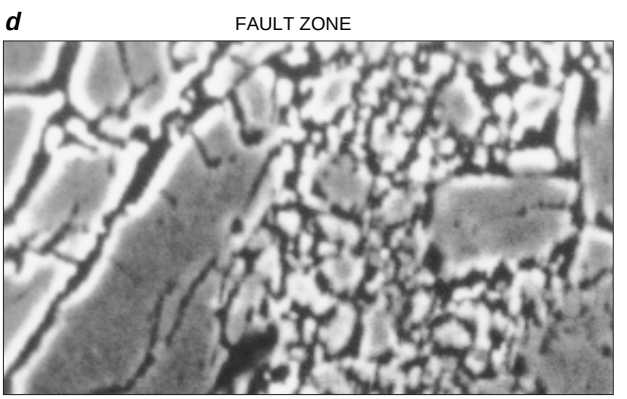
Brittle failure and anticrack failure, the mechanisms that account for shallow and deep earthquakes, respectively, share many characteristics. Both processes involve the development of microscopic features (*a*) that cooperatively form a fault and allow for movement on it.

As brittle rocks experience mounting stress, microcracks open parallel to the direction from which they are compressed. In rocks found more than 300 kilometers deep, pressure prohibits this dilation. Instead microanticracks (lenses filled with fine-grained, dense spinel-phase olivine) form perpendicular to the direction from which these rocks are compressed.

In each mechanism, at some critical point in time, these microfeatures link up and create a fault (*b*). Movements along the fault then relieve stress (*c*). In brittle failure, the fault is an open fracture. In anticrack failure, the fault contains fine-grained spinel-phase olivine. This material is superplastic—that is, the crystals can readily move past one another, enabling the fault to slide. Because the fault need not dilate to slip in this way, pressure does not restrict the process.

Brittle fault zones contain angular crystals showing a fractal size distribution. In anticrack fault zones, rounded olivine fragments are embedded in extremely fine grained spinel-phase olivine (*d*).





is carried deep into the upper mantle, where the spinel crystal structure is stable. In particular, this mechanism cannot account for earthquakes shallower than approximately 300 kilometers, where olivine is still stable. Normal brittle fracture, though, cannot explain earthquakes deeper than 70 kilometers. What transpires in between these depth regions? Other recent experiments have neatly provided the explanation for such intermediate-focus earthquakes.

Working at the University of California at Berkeley, Charles Meade (now at the Carnegie Institution of Washington) and Raymond Jeanloz showed that the hydrous mineral serpentine (which forms when olivine reacts with water at low temperatures and pressures) emits acoustic energy when dehydrated at very high pressure under stress. C. Barry Raleigh, now at the University of Hawaii at Manoa, and Mervyn S. Paterson of the Australian National University in Canberra demonstrated dehydration-induced faulting of serpentine in the 1960s, but at low pressure. Meade and Jeanloz's experiments were similar but were carried out instead on sand-grain-size specimens of serpentine in a diamond-anvil cell. The serpentine emitted acoustic energy when it was heated and dehydrated under pressures equivalent to that found 300 kilometers deep in the earth. We can understand this process in terms of the anatomy of brittle fracture. The pressure of the water produced by dehydration pushes open microcracks against the high applied pressure, thereby allowing for brittle failure.

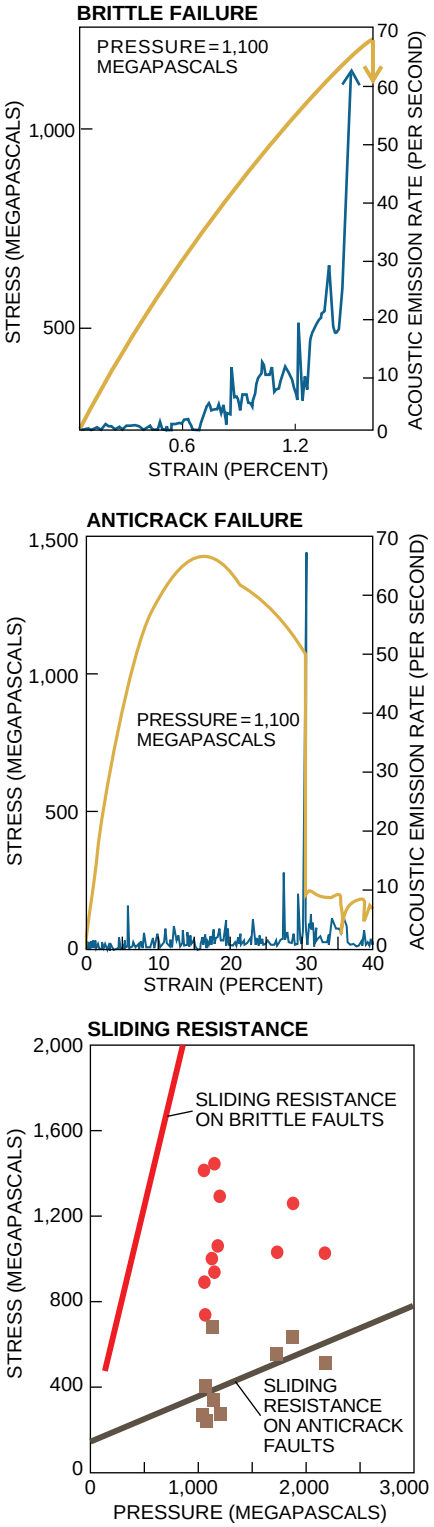
We know from a variety of geophysical and geologic observations that olivine in the uppermost mantle (just below the oceanic crust) becomes partially hydrated as it journeys from an ocean ridge to an ocean trench. Thus, shallow regions in the lithosphere contain the hydrous phases that enable this mechanism to work. The declining frequency of earthquakes in subduction zones down to 300 kilometers most probably represents the progressive exhaustion of the mechanism as the oceanic lithosphere gradually warms up and dehydrates, heated by the surrounding mantle. At about 300 kilometers, anticrack faulting becomes possible, causing an increase in earthquakes there.

The anticrack faulting mechanism provides an explanation for how and why earthquakes extend to great depth in the earth. Can this mechanism also explain why they suddenly stop? As mentioned earlier, the decomposition of spinel into two denser phases occurs at approximately 700 kilometers deep in subduction zones. This decomposition reaction is endothermic (it requires the addition of heat to proceed). In contrast, the transformation from olivine to spinel is exothermic (heat is released during the reaction). If we were correct in our original assumption that a thermal runaway must occur to introduce a faulting instability, then an endothermic reaction should be incapable of producing such an instability.

To test this possibility, just this year, my colleague Yi Zhou and I conducted a set of experiments on $CdTiO_3$, a composition that undergoes an endothermic densification transformation. Deformation of the low-pressure phase under conditions for which the high-pressure phase is stable proceeded uneventfully; neither anticracks nor faulting was observed. This powerful test supports both the anticrack model and our reasoning as to why earthquakes cease at the upper pressure limit of spinel stability. Not only is the breakdown reaction endothermic, but it also requires the unmixing of atoms to produce two crystal structures from one. Such a transformation would further inhibit any potential faulting instability.

In summary, the depth distribution of earthquakes and the experimental results lead naturally to the following model. Normal brittle fracture accounts for shallow earthquakes. Because pressure inhibits this mechanism, in most parts of the world earthquakes take place only down to 20 to 30 kilometers below the earth's surface. In subduction zones, partially hydrated oceanic crust and mantle sinks downward and is slowly heated. The water-bearing minerals are dehydrated and, in the process, make fluid-assisted faulting possible. The exponential decrease in earthquake frequency down to 300 kilometers reflects the progressive heating and dehydration of the subducting slab.

The interior of the slab remains sufficiently cold so that the olivine of the subducting mantle cannot transform to the spinel phase when it leaves the olivine stability field at approximately 300 kilometers. At the margins of this cold interior region, the temperature slowly increases. The metastable olivine heats to the critical temperature at which anticrack faulting takes place. In the coldest subduction zones, the wedge of metastable olivine extends down approximately 700 kilometers, where it then decomposes into the two very dense phases of the lower mantle. After



ACOUSTIC EMISSIONS (*blue*) occur when a fault suddenly slips, radiating energy that relieves stress (*yellow*). Brittle failure, the mechanism responsible for shallow earthquakes, emits noise before and during movement on a fault (*top*). In contrast, anticrack failure, the mechanism behind deep earthquakes, emits acoustic energy only when faulting takes place (*middle*). At higher pressures, brittle faulting can occur only under much greater stress (*bottom*). For this reason, brittle failure cannot explain deep-focus earthquakes. Pressure does not likewise inhibit anticrack failure.

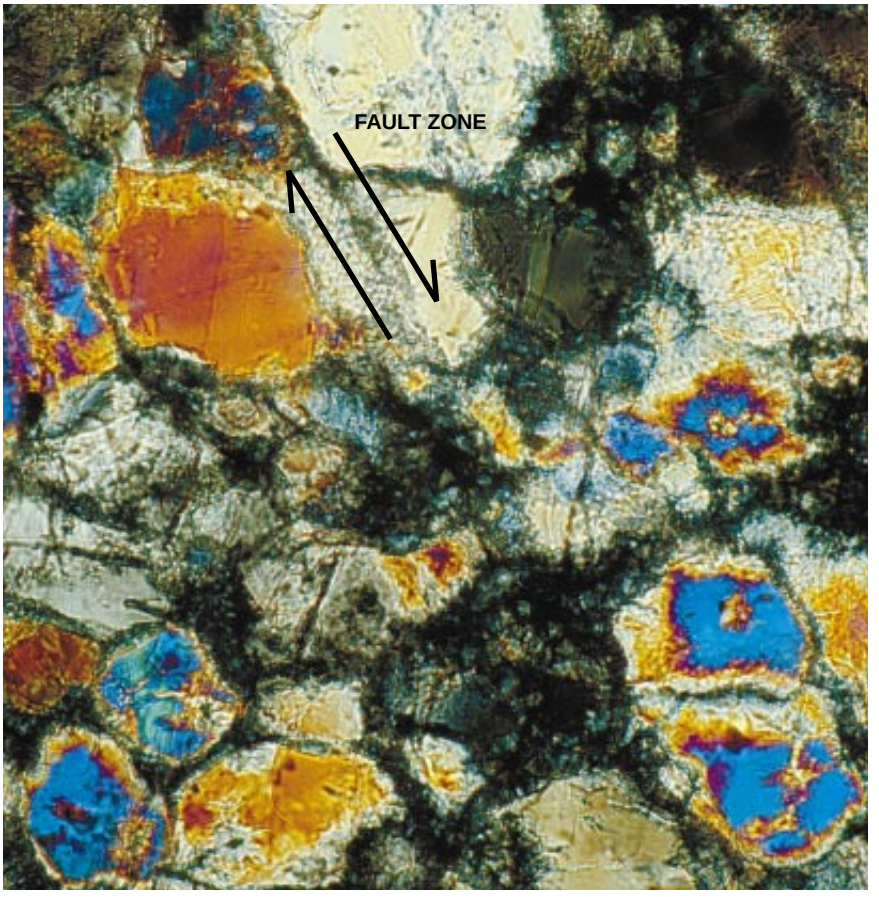
this deep reaction, all earthquakes stop.

The model automatically predicts certain properties for the seismic signals generated in the mantle during intermediate- and deep-focus earthquakes and for the changes in seismic velocity within subducting slabs. First, the seismic signal of these earthquakes should be highly similar to those of shallow earthquakes. In particular, their signals should be consistent with shearing motions on a fault. Indeed, this seems to be the case. Although seismologists have searched for the past three decades, they have found no unambiguous instance of a deep earthquake having a strong implosive component.

Moreover, the seismic velocity of the cold interior of subducting plates should be significantly slower if metastable olivine is present than if the reaction has already run and produced the denser polymorphs. Only Japan experiences a sufficient amount of deep earthquakes and has enough seismic stations to attempt to distinguish between these two possibilities. In 1992 Takashi Iidaka and Daisuke Suetsugu of the University of Tokyo modeled both possibilities for the descending slab underneath Japan and found the telltale slow velocity of a metastable olivine wedge.

If, as we propose, a critical temperature controls the anticrack faulting instability, the faulting will be concentrated at the interface between the metastable olivine wedge and the surrounding, already transformed carapace. If sufficient stress exists on both margins of the wedge, double zones of earthquakes could develop. Two sets of seismologists, one led by Douglas A. Wiens of Washington University and the other by Iidaka, have discovered such double zones during the past year. Iidaka's team found the double zone in the slab that they previously proposed must contain a metastable olivine wedge.

In addition, if there is a change in the fundamental mechanism responsible for earthquakes at 300 to 400 kilometers, one might expect some aspects of the seismic signals generated at such depths to be different from those generated by shallower events. Until very recently, all attempts to identify such distinctions had failed. Still, Heidi Houston and Quentin Williams of the University of California at Santa Cruz recently reported that many deep events seem to start up significantly faster than do intermediate events. Further, Houston and John E. Vidale of the U.S. Geological Survey have now determined that the entire rupture time of such earthquakes is only about half that of shallower ones. Other recent work has shown that young, warm subduction zones experience earthquakes only down to 300 to 400 kilometers; all deep earthquakes are confined to older, colder subduction zones, where metastable olivine is likely to persist to great depths.



FAULT ZONE, produced by deforming a sample of mantle olivine, extends from the top left corner to the bottom right corner of the micrograph. An offset olivine crystal (*white*) appears on this fault in the top half of the image. The anticracks (*yellow lenses in blue crystal*) that generate the fault zone form perpendicular to the direction from which the sample is compressed. They grow throughout the material parallel to this direction.

The laboratory results therefore explain how earthquakes can happen at very high pressures. The composite model advanced here, in which intermediate-focus earthquakes occur by fluid-assisted faulting and deep events by anticrack faulting, is highly consistent with our current understanding of subduction zones. In the field, seismologists continue to characterize subduction zones and their earthquakes, and geophysicists are developing quantitative thermal models that incorporate both experimental and seismic information. Questions remain, but the essential paradox behind deep earthquakes has been resolved.

FURTHER READING

DEEP EARTHQUAKES. Cliff Frohlich in *Scientific American,* Vol. 260, No. 1, pages 48–55; January 1989.

A NEW SELF-ORGANIZING MECHANISM FOR DEEP-FOCUS EARTHQUAKES. H. W. Green II and P. C. Burnley in *Nature,* Vol. 341, No. 6244, pages 733–737; October 26, 1989.

ANTICRACK-ASSOCIATED FAULTING AT VERY HIGH PRESSURE IN NATURAL OLIVINE. Harry W. Green II, Thomas E. Young, David Walker and Christopher H. Scholz in *Nature,* Vol. 348, No. 6303, pages 720–722; December 20–27, 1990.

SEISMOLOGICAL EVIDENCE FOR METASTABLE OLIVINE INSIDE A SUBDUCTING SLAB. Takashi Iidaka and Daisuke Suetsugu in *Nature,* Vol. 356, No. 6370, pages 593–595; April 16, 1992.

EVIDENCE FOR TRANSFORMATIONAL FAULTING FROM A DEEP DOUBLE SEISMIC ZONE IN TONGA. Douglas A. Wiens, Jeffrey J. McGuire and Patrick J. Shore in *Nature,* Vol. 364, No. 6440, pages 790–793; August 26, 1993.

Privatizing Public Research

With the end of the cold war, national defense has given way to international competitiveness as the theme for federal support of research. As it now stands, the idea will probably not work well

by Linda R. Cohen and Roger G. Noll

Political support for federally funded research and development is apparently beginning to unravel. Adjusted for inflation, the government's R&D expenditures have fallen 7 percent since 1988. Spending on R&D in the private sector is still increasing, but its growth has fallen below the rate at which output is increasing. Investment in research is not keeping pace with the economy.

In large part, these trends reflect a fundamental change in the rationale for federal research expenditures. From the beginning of World War II through the late 1980s, national security concerns dominated R&D policy. More than half the federal R&D budget was devoted to defense technology, and much of the rest—including fundamental research in mathematics and the physical sciences—received support because of its potential relevance to national security. The end of the cold war weakened this justification for federal research policies.

During the past decade, government officials have sought new goals for their research dollars. The most important emerging theme in their programs is international competitiveness: the federal government should support R&D to increase American industrial productivity, thereby helping industry in global economic competition.

We believe the new competitiveness rationale will not succeed in reinvigorating the national R&D effort. First, competitiveness is not a politically powerful substitute for the cold war in forging a durable, bipartisan coalition for supporting R&D at the generous levels typical of past decades. Second, the methods for implementing the new programs are shaped by political necessity and so are likely to undermine the economic performance of the programs. Eventually, that will further reduce the political support for the programs.

Historically, the desire to help an industry increase productivity has always played some role in R&D policy. Federal subsidies for commercially relevant R&D are more than 100 years old, having supported the development of the telegraph and hybrid seeds in the 19th century. Nevertheless, commercial programs did not become a significant component of federal R&D support until World War II. Even then, these programs were almost exclusively targeted at defense-related technologies.

Not until the 1960s did the federal government undertake a broad array of research programs for primarily civilian purposes. These programs were not part of a coherent plan for fostering national economic growth. Instead they were a series of largely unrelated responses to much narrower public issues that gave rise to new "missions" for federal agencies. Examples were the war on cancer, the drive to develop environmentally benign technologies and attempts to find an effective response to the rise of the worldwide oil cartel.

Despite these initiatives, most federal R&D dollars were still spent on defense or on fundamental knowledge directly relevant to defense. Most proposals to broaden the base of these programs—either by adopting a comprehensive commercial R&D policy or by adding more industries to the list of those receiving support—were defeated. In contrast, the current approach to R&D is essentially economy-wide. Its appeal rests on the argument that it can help U.S. industry boost productivity and reclaim dominance in international markets. Almost any industry is a possible target for support.

The competitiveness theme has caused two major changes in how federal R&D programs are formulated and managed. One change is greater privatization of the selection and results of research projects. Privatization is perhaps most clearly evident in the extent to which the new programs assign to private industry both responsibility for decisions about technical choices in the projects and essentially all the intellectual property rights. The other change is increased collaboration among American firms and research organizations.

The most extreme example of these changes is in the open competitions held by the National Institutes of Standards and Technology for its Advanced Technology Program (ATP). Any firm or group of firms, in any industry, can submit a proposal for partial federal funding for a technology development project. Proposals are evaluated by criteria that include potential commercial success, a feasible commercialization and marketing strategy, technical interest, inability to obtain complete private backing for the project and the likelihood of broad applications. These projects need not relate to any specific government mission. Indeed, the program avoids projects intended to provide technology for use by a government agen-

LINDA R. COHEN and ROGER G. NOLL both study issues at the boundary of economics, law and political science. Cohen is currently a visiting professor in law and economics at the University of Southern California and the California Institute of Technology. She is on leave from the University of California, Irvine, where she is associate professor of economics. Cohen holds a bachelor's degree in mathematics from the University of California, Berkeley, and a Ph.D. in social science from Caltech. Noll, professor of public policy in the department of economics at Stanford University, received an undergraduate degree in mathematics from Caltech. His doctorate is in economics from Harvard University.

WAFER TECHNOLOGY is used in developing improved flat-panel displays at a Kopin Corporation plant in Massachusetts under a contract from the Advanced Research Projects Agency. The work represents the new theme of federal support for R&D—helping U.S. firms in international competitiveness.

cy such as the Department of Defense.

Most of the funds for carrying out the new competitiveness strategy still go to programs that superficially retain more of a public-sector focus. The Technology Reinvestment Program, administered by the Advanced Research Projects Agency in the Department of Defense, has an annual budget of more than half a billion dollars. Its goal is to support projects that will allow firms to rely on commercial markets and profits while developing technologies useful for defense. In addition, several large industry-specific programs have been established, including Sematech (for semiconductor manufacturing technology), the flat-panel display program, the Clean Coal Technology Program, the National Aerospace Plane Consortium and the Clean Car Initiative. All of them receive substantial federal monies.

To help justify their own continued existence, the federally funded national laboratories are also seeking to conduct joint research with companies. These efforts are called Cooperative Research and Development Agreements (CRADAs). Like the ATP projects, CRADAs are available for all industries and need no connection to any government program.

Despite their contributions from the public sector, these undertakings all diverge from traditional Department of Defense and Department of Energy R&D programs. Commercial technology development is a primary goal rather than an unsought (but welcome) spin-off from carrying out a government mission. Each venture relies on the private participants to propose and manage the projects. Each requires that property rights belong to the private participants rather than to the government sponsor or partner. All the activities require private enterprises to share in the costs. As with the ATP, these programs involve an unusual amount of proprietary information. Project proposals, for example, are routinely exempted from the Freedom of Information Act and are reviewed exclusively by government agencies.

The principle that economic growth is enhanced by new technology has a firm foundation in theoretical and empirical research in economics and has been investigated by several distinguished economists, among them Nobel Laureate Robert M. Solow of the Massachusetts Institute of Technology, the late Edward Dennison of the Brookings Institution, Moses Abramovitz of Stanford University, Edwin Mansfield of the University of Pennsylvania, Richard Nelson of Columbia University and Zvi Griliches and Fredric M. Scherer of Harvard University. The main conclusions from their work are that more than half the historical growth in per capita income in the U.S. is attributable to advances in technology and that the total economic return on investment in R&D is several times as high as that for other forms of investment.

That R&D can enhance the nation's economic welfare is not, by itself, sufficient reason to justify a prominent role for the federal government in financing it. Economists have developed a further rationale for government subsidies. Their consensus is that most of the benefits of innovation accrue not to innovators but to consumers through products that are better or less expensive, or both. Because the benefits of technological progress are broadly shared, innovators lack the financial incentive to improve technologies as much as is socially desirable. Therefore, the government can improve the performance of the economy by adopting policies that facilitate and increase investments in research.

In principle, government can solve the problem of underinvestment in R&D in two ways. The first, which political conservatives tend to emphasize, is to promote the ability of innovators to obtain higher profits. In the past, the most important policy for augmenting the profitability of innovation has been to strengthen intellectual property rights: patents, copyrights and legal protection of trade secrets.

This approach has two significant drawbacks. First, it creates higher profits through the establishment of monopolies, which are inefficient. Second, any form of protection for intellectual property limits the diffusion of the research results. Research often has applications in a variety of products and industries; the possible benefits of a discovery can be realized only if people other than the discoverers have the opportunity and incentive to apply those findings in new ways.

The other approach, which liberals tend to favor, is for the government to pay for R&D through targeted programs. In this case, the government selects specific technologies and projects; it then either subsidizes them in the private sector or undertakes them in government research laboratories. This approach, too, has drawbacks. If the goal of the program is to encourage commercial successes, the government is not likely to pick the best projects. Furthermore, monitoring public research projects to assure that private contractors are putting forth their best efforts is notoriously difficult.

The root of the monitoring problem is uncertainty about both the costs and the results of R&D projects. By the very nature of R&D, costs and results are imperfectly known—otherwise the research would not have to be done in the first place. Consequently, the government faces difficulties in specifying realistic technical approaches and objectives. Moreover, early work on a project is likely to generate information that moves the government to change the details of the project. Contracts are therefore often based on a cost-reimbursement formula. Unfortunately, such contracts are well known for their tendency to produce cost overruns.

The government's traditional safeguard against companies that take advantage of cost and performance uncertainties is to impose rigorous cost-accounting and auditing requirements on R&D contractors in a herculean effort to find waste, fraud and lax management. This monitoring system applies to most of the new programs, and it is far more elaborate, costly and inflexible than the monitoring systems that private organizations employ for their own research. As a result, R&D done under federal contract is inherently more expensive and less effective than R&D done by an organization using its own funds. In fact, government monitoring methods are so burdensome that many federal contractors separate their federal and private work so that they can use more flexible, cheaper methods for managing their private R&D.

In the past, a political consensus for federal R&D was achieved by pursuing both approaches and by treating them as substitutes. Technologies in which private industry could hold a reasonably secure intellectual property right were expected to be backed by business, whereas the new knowledge emanating from government-supported R&D was to be nonproprietary and widely disseminated. Even in defense technologies, where national security considerations led officials to keep many research results secret, the government frequently called on contractors to produce technologies that would be further developed by other contractors. The government also encouraged defense companies to disseminate their particular technical knowledge through subcontracting. It welcomed the commercial adoption of technologies that were not closely related to highly confidential defense products, such as advances in computers, microelectronics and telecommunications.

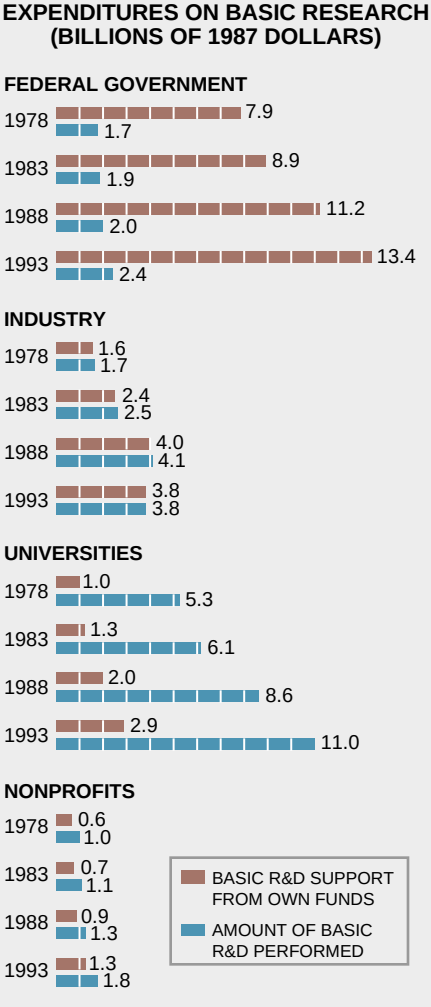
The new theme preserves both approaches, thereby appealing to conservatives and liberals, but rejects the notion that the approaches are substitutes. Proponents argue that privatization is necessary for competitiveness (that is, for limiting the benefits of the programs to domestic firms) and for giving research organizations sufficient incentive to bring new technical knowledge to commercial practice. They buttress their case by pointing to the similar approach to R&D undertaken in Japan (for computers and microelectronics) and Europe (for Airbus). Those governments established and subsidized collaborative, proprietary R&D projects in the private sector that have led to important gains in the international competitiveness of their domestic industries.

Control of intellectual property from government money has become increasingly private since the passage in 1980 of the Bayh-Dole Act, which allows both companies and universities to hold rights to inventions supported by government contracts and grants. The ATP program also privatizes the results of its federally subsidized projects. Its authorizing statute explicitly

R&D Funding at a Glance

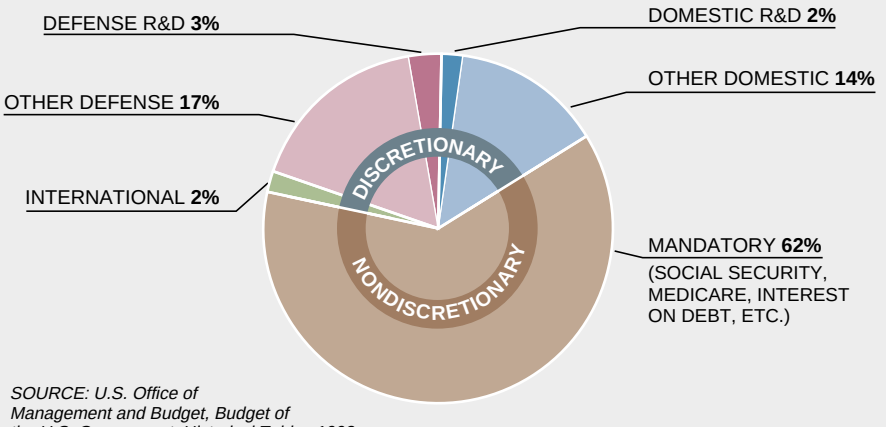
Federal funding of research and development has shifted in emphasis since the end of the cold war. Until then, almost all the money supported defense technologies or programs potentially relevant to national security, such as research in mathematics and the physical sciences. Now the emphasis is on international competition, and private enterprise plays a much larger role.

Funds dedicated to R&D made up about 5 percent of the federal budget in fiscal 1993 (*top right*). Only part of the government's allocation for basic research is spent at national laboratories; the rest supports research in industry, universities and nonprofits (*below*). Federal spending on all university research has risen, but the share of university research funding that comes from the government has declined (*middle right*). Nongovernmental sources of funding have therefore become increasingly important to R&D support (*bottom right*).



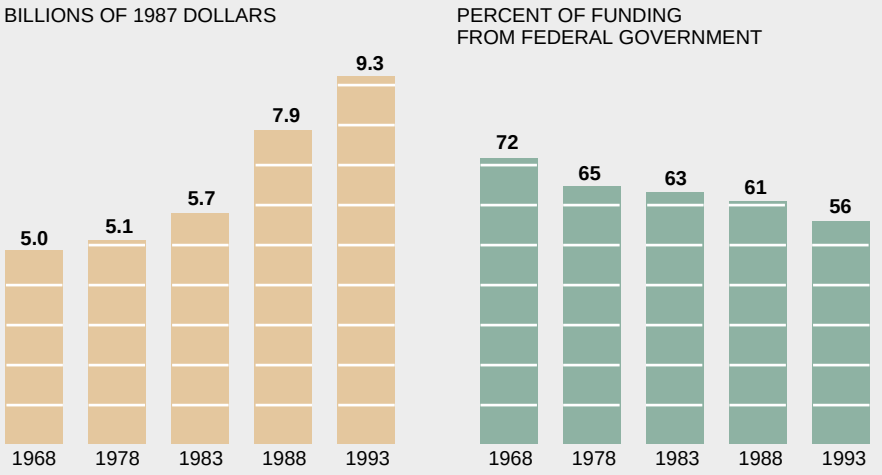
SOURCE: U.S. National Science Foundation, Science and Engineering Indicators 1993

DISTRIBUTION OF FEDERAL SPENDING (1993)



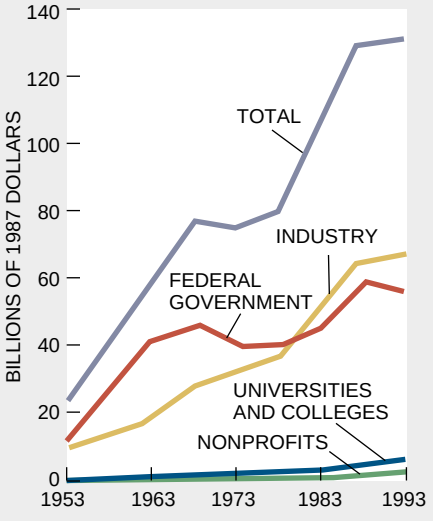
SOURCE: U.S. Office of Management and Budget, Budget of the U.S. Government: Historical Tables 1993

FEDERAL FUNDING OF UNIVERSITY RESEARCH

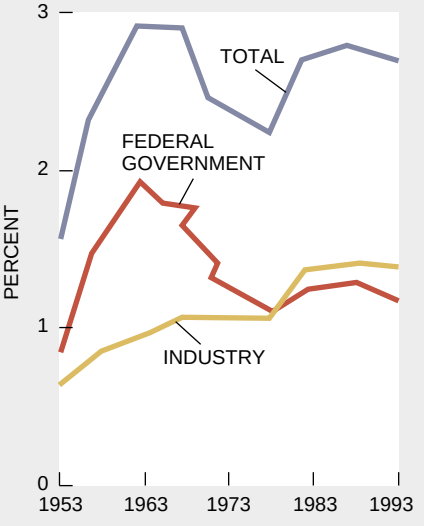


SOURCE: U.S. National Science Foundation, Science and Engineering Indicators 1993

FUNDING FOR R&D, BY SOURCE OF SUPPORT



R&D AS A SHARE OF GROSS DOMESTIC PRODUCT, BY SOURCE OF SUPPORT



SOURCE: U.S. National Science Foundation, National Patterns of R&D Resources

requires that any patents from the work must belong to profit-making firms, thereby excluding universities.

An economic consequence of privatizing knowledge is that it is more effective in encouraging the development of products than in generating scientific advances. A company usually cannot capture as much of the economic value of an advance in fundamental knowledge as of an advance in a product or production technique. The implication of this observation can be seen most dramatically at universities. Historically, the U.S. government has contributed to university research and encouraged the open and free dissemination of the results. That policy—both directly and through the movement of students to industry or other universities—has offered significant benefits. Recent studies, pioneered by Mansfield, show that the economic payoff from fundamental research is higher in the U.S. than in other nations.

The policy of assigning greater project control and management to the private sector is likely to shift spending toward the kind of research done by industry in the past. Most basic research has been financed by the government, through generous grants to research universities and also through the basic-research component of the work in federal laboratories. About 20 percent of the federal R&D budget is expended on basic research, compared with less than 6 percent in industry. The peak year for federal R&D support was 1988, when the government accounted for 46 percent of the total U.S. R&D budget but 62 percent of basic research.

Meanwhile industry, which paid for half of all R&D, paid for only 22 percent of basic research. As universities have come to rely more on industry for research support, the character of their work has shifted accordingly. In the past 20 years, the share of university research that is basic has fallen from 77 to 64 percent. Most of the decline has occurred in the past few years.

Partly to encourage broadly applicable R&D, the new government programs propose that projects be undertaken by consortia or joint ventures of firms, perhaps together with universities and other research institutions. Initially, the ATP favored the formation of consortia that would conduct "generic, precompetitive" research, which the member companies could later apply to the development of competing products. Thus, the policy has aimed to foster those research activities that the private sector has traditionally given the shortest financial shrift. In the past decade, the federal government has also tried to

stimulate joint research by relaxing certain antitrust standards.

Research collaborations can be beneficial. If a firm is unlikely to invest in a research activity because it cannot capture the benefits, a joint venture may make the effort worthwhile. An example is research that expands the technological base of an industry but allows each company to develop its own products and production methods. Collaboration can also enable companies to benefit from synergies, in that each firm may bring different expertise to the venture.

Collaborative research raises a dilemma, however. Although in theory it can encourage research investments, it can create cartels. Obtaining the maximum security for intellectual property rights requires the elimination of market competition, which in turn necessitates that a cartel take control of the domestic industry and that barriers be erected against the import of the cartelized products.

The postwar history of Japan illustrates this inherent problem. By facilitating the formation of cartels, Japan produced a system that ranks first in the world in its fraction of gross domestic product (GDP) invested in R&D. It has a higher rate of sustained growth than any other advanced, industrial economy, and it has consistently low unemployment. Moreover, in Japan a far lower proportion of R&D is financed by government than in any of the world's other leading economies, and defense plays virtually no role in motivating public or private R&D.

The other side of Japan's remarkable performance is that the ordinary Japanese citizen has a standard of living far below that of citizens in other nations with approximately the same per capita GDP. Although on average Japanese employees work substantially more hours per year than American workers, the real purchasing power of the average annual take-home pay in Japan is only about 75 percent of that in America—a noteworthy fact, considering that the real value of wages for average American workers has not increased for 20 years.

Furthermore, most Japanese are not pleased with their economic system. The recent political upheaval in Japan was sparked by scandals among members of the ruling party, but it has been brewing for more than a decade. It is rooted in the ordinary Japanese citizen's dissatisfaction with being unable to attain a standard of living roughly equal to that in North America and western Europe. The key lesson from the experience of Japan is that although policies providing extremely high financial incentives for private investment will produce rapid economic growth, the economic benefits of that growth will not be widely shared.

The implications of collaboration for the rate of innovation are thus complex. They depend on the details of the work done and the role of the domestic industry in the international economic system. R&D collaboration among domestic competitors is unlikely to have the undesirable effects of a domestic cartel in either of two cases. If world trade is free and several nations are efficient producers, even a complete domestic cartel in an industry will be merely another competitor in the global market. The collaboration might enable the domestic industry to achieve economies of scale or scope in R&D. A collaboration is also likely to be beneficial if its scope is limited to expanding the technological base of the domestic industry: each firm in the industry can then make use of that technology to develop its own proprietary products.

Although a logical argument could be made that a consortium could efficiently engage in research devoted to enhancing an industry's underlying technology, firms in competitive industries most often focus on technology that can be more easily protected. The most widely hailed successes of Sematech, for example, have been in developing semiconductor manufacturing equipment. Members of the consortium have then made advantageous use of the new machines.

Many of the horizontal joint ventures of the ATP follow similar strategies. The Advanced Display Manufacturers of America Research Consortium has been involved in the development of automated inspection and repair equipment that members use in producing a variety of flat-panel displays. The Rapid Response Manufacturing project (a consortium led by the National Center for Manufacturing Sciences) brings together a group of firms specializing in computer-aided design and four manufacturing firms to develop software that demonstrates the design of crankshafts, steering columns, microwave antennas and aircraft engine covers.

These efforts are all "generic" from the viewpoint of the member firms. The projects deal with difficult and fascinating technical issues and, by drawing in firms from diverse industries, may well result in valuable synergies and the wider diffusion of results than could be expected from an effort by a single company. Yet virtually all these projects seek to pursue technologies that can be protected by patents and limited, at least in part, to use by members of the consortium.

Indeed, technology-base activities that cannot be protected by patents poorly address the competitiveness goal of the new programs. Foreign competitors are as likely as U.S. firms to benefit from such research efforts. And if domestic

EUROPEAN AIRBUS was the product of a consortium subsidized by several European governments. Such collaborations represent one governmental approach to the goal of improving the international competitiveness of domestic industries.

companies outside the consortium separately apply the new technology base to the development of competing products, the profitability of the R&D work diminishes. Thus, such activities are unlikely to generate much enthusiasm from business—unless, as was the case in Japan, they are accompanied by a domestic production cartel and strong trade barriers.

The industry-wide centralization of applied R&D therefore confronts domestic consumers with a Hobson's choice. If the venture makes U.S. industry more productive than its foreign competitors, the domestic industry will retain most of the benefits of its expanded productivity by cartelizing the domestic market. If the venture fails to make U.S. industry more competitive, the domestic industry will lose market share to foreigners, leading to the imposition of import restrictions. Once again, a domestic cartel emerges, but in this case one that is inefficient as well as monopolistic. In either case, the main effect of centralized R&D is to transfer wealth to members of a domestic cartel, not to promote the economic welfare of most citizens.

COMMUNICATIONS SATELLITE hovers over the earth shortly after its launch from the space shuttle *Endeavor*. Such products represent the dominant theme of federal support for R&D until the end of the cold war: national security.

If the centralization of applications research for an entire domestic industry is dangerous, what is the alternative? One possibility is to support projects undertaken by only part of the domestic industry. That approach has been taken by the ATP, which even supports separate competing proposals by different groups of firms that are pursuing advances aimed at the same products. It is also the method proposed by the Department of Defense for the flat-panel display program. This strategy can be effective because it can preserve competition while subsidizing proprietary research.

In practice, such programs have run into an obstacle: an inverse relation between performance and political viability. When a project becomes successful, outside firms perceive it as unfair on the grounds that the government is interfering with the success or failure of companies in the industry. For example, Cray Research and two Department of Energy laboratories entered into a highly publicized $52-million CRADA to develop supercomputing technology. It was unceremoniously shelved after complaints from other supercomputer firms that the government's subsidy gave Cray a competitive advantage.

More often, such programs are not very successful. To some degree, failures should be expected because the outcome of an R&D effort is inherently unpredictable. Unfortunately, technical and economic failure seldom lead to the timely demise of a major project. Government officials, unlike corporate executives, must be sensitive to the effects that canceling a project will have on employment. In short, the government has difficulty completing successful projects and difficulty cutting its losses on failures.

Some planners have argued that requiring an industry to share in the costs of R&D would provide a motivation to discontinue funding for failing projects. That approach may work for small projects, but a big one—one that generates substantial employment and accounts for a large share of some company's business—becomes too politically important for the government to abandon it easily. Congress is more likely to revise the program so that the government assumes a greater financial burden. It was the cost-sharing requirements, not the projects, that were abandoned when industry was unwilling to shoulder its share of the cost overruns for the Clinch River Breeder Reactor and the supersonic transport airplane.

Our conclusion is that the U.S. has not yet found a politically workable and economically attractive means of encouraging technological progress. Both economic research on R&D and the historical experience with government programs indicate that the most effective solution would be a combination of policies. The government would take a directing role in subsidizing both fundamental research and research aimed at broadening the technology base. It would make the results of that research widely available rather than proprietary. For applications research, the government can probably encourage development and avoid political pitfalls only if it follows a completely hands-off policy (say, by providing differential tax treatment for research investment) or if it concentrates on small schemes in carefully selected industries. The latter strategy is unlikely to have any wide-ranging impact on economic performance.

Unfortunately, this approach entails significant political liabilities. Consumers would benefit the most—but most of the political support for R&D comes from industry, not consumers. And because the entities that conduct technology-base research cannot usually keep the results for themselves, U.S. firms would not be the only beneficiaries. In the current political lexicon, a "competitiveness" strategy is one with policies that focus on short-term commercial products and exclude foreign companies from sharing in the benefits. That interpretation is fundamentally at odds with the broader concept of promoting innovation.

The breadth and intensity of the federal R&D programs that this nation has enjoyed in the past have richly contributed to the growth of its economy. Maintaining that support, however, hinges on building a strong, stable consensus for it. Competitiveness cannot replace national security as the basis for that consensus.

FURTHER READING

THE TECHNOLOGY PORK BARREL. Linda R. Cohen and Roger G. Noll. Brookings Institution, 1991.

BEYOND SPINOFF: MILITARY AND COMMERCIAL TECHNOLOGIES IN A CHANGING WORLD. John A. Alic, Lewis M. Branscomb, Harvey Brooks, Ashton B. Carter and Gerald L. Epstein. Harvard Business School Press, 1992.

THE GOVERNMENT ROLE IN CIVILIAN TECHNOLOGY: BUILDING A NEW ALLIANCE. Committee on Science, Engineering, and Public Policy. National Academy Press, 1992.

THE RISE AND FALL OF AMERICAN TECHNOLOGICAL LEADERSHIP: THE POSTWAR ERA IN HISTORICAL PERSPECTIVE. Richard R. Nelson and Gavin Wright in *Journal of Economic Literature,* Vol. 30, No. 4, pages 1931–1964; December 6, 1992.

EMPOWERING TECHNOLOGY: IMPLEMENTING A U.S. STRATEGY. Lewis M. Branscomb et al. MIT Press, 1993.

The Scientific Importance of Napoleon's Egyptian Campaign

Bonaparte's invasion of Egypt brought French scientists and engineers to the Nile. Their work, in turn, brought the splendors of the Nile to Europe

by Charles C. Gillispie

On the first of July 1798, an armada of 400 ships appeared off the coast at Alexandria. By the end of the day, longboats had put ashore an army of 36,000 men under the command of Napoleon Bonaparte. Meeting with no resistance, he immediately marched his troops, sweltering in their woolen Alpine uniforms, through the oven of the desert to rout the Mameluke rulers of Egypt in the Battle of the Pyramids on July 21. Ten days later Admiral Horatio Nelson destroyed the French fleet, marooning the expeditionary force in the land it was to control and explore for the next three years.

Bonaparte abandoned his soldiers a year later, slipping through the British blockade and returning to France to seize power in the coup d'état of November 9, 1799. Among the handful of retainers he took with him were Gaspard Monge and Claude-Louis Berthollet, the leading members of the first scientific task force to accompany a military expedition. Their colleagues from the Commission of Science and Arts—a group of 151 scientists, engineers, medical men and a few scholars—were left behind with the army. The elite among them were elected to the Institute of Egypt, founded on Bonaparte's initiative as a colonial adaptation of the Institute of France. Serving as permanent secretary throughout the occupation was Jean-Baptiste Fourier, who had yet to invent the analysis that bears his name.

The most famous discovery of the expedition remains the Rosetta stone, now in the British Museum. The French surrendered it with immense reluctance to the British forces that expelled them from Egypt late in 1801. The commission of technical experts accomplished a great deal else of scientific interest in the land of the pharaohs. A compilation of monumental dimensions contains the record of their archaeological surveys; of their research into the physical and chemical phenomena as well as the natural history peculiar to the region;

CHARLES C. GILLISPIE is a historian of science who is known as editor of the *Dictionary of Scientific Biography.* After completing undergraduate work in chemistry at Wesleyan University, he received his doctoral degree in history from Harvard University. Since 1947 Gillispie has taught at Princeton University, where he is now professor emeritus and where he established the program in history of science in the 1960s.

and of their inquiries into the sociology of an exotic country.

La Description de l'Égypte, printed between 1809 and 1828, required a specially designed piece of mahogany furniture to house it. Ten folio volumes of plates measuring 20 by 26 inches and two atlases, 26 by 40 inches each, contain 837 copper engravings (50 in color and many with multiple illustrations). A third atlas consists of a topographical chart of Egypt and the Holy Land in 47 sheets. Nine accompanying volumes of text dwarf any modern encyclopedia. They comprise approximately 7,000 pages of memoirs, description and commentary. The whole is divided into three parts: ancient Egypt, modern Egypt and natural history.

The archaeological plates of part one, which make up just over half the total wealth of illustrations, created the first modern vision of Egyptian antiquity. It was in contemplating the sweep and detail of those enormous and powerful engravings that Europeans took the measure of the valley of the Nile. Earlier Western awareness of the land of the pharaohs had consisted mainly of hearsay about the scale and orientation of the Pyramids and the mystery of the Sphinx. Of Upper Egypt nothing was known beyond the odd traveler's tale of some giant arm, say of Shelley's *Ozymandias,* thrusting out of the sand. An account by the artist Vivant Denon, who accompanied the soldiers in their campaign up the Nile, conveys the impact of rounding a bend in the river and coming on the temples of Karnak and Luxor amid the ruins of Thebes: "The whole army, suddenly and with one accord, stood in amazement…and clapped their hands with delight."

The contributors to *La Description de l'Égypte* captured on paper all the monuments, starting in the south at the Isle of Philae. Drawing, measuring and excavating along the way, they moved downstream through Kom Ombo and Edfu—close by the Nile on the right and left banks, respectively—and past Esna, slightly set back from the river to the west. Their party paused longest amid the vast array of Thebes, awed by Medinet Habu, the Ramasseum and the colossal statues of Memnon, behind which lie the tombs in the Valley of the Kings and facing which loom the enormous piles of Luxor and Karnak across the Nile. Downstream they came upon the architectural and artistic climax of Dendara. After recording these masterpieces, the group continued north to Memphis and the Pyramids at Giza.

Each site is depicted in a sequence of eight to 10 plates, beginning with topography. Next comes a panorama of the structure in its condition at the time, choked with sand, columns cracked and tumbled, ramparts crumbling, the overall majesty somehow enhanced. Architectural drawings follow, providing ground plans, sections and elevations. Several sheets then depict architectural detail, bas-reliefs and other sculptures as well as surfaces covered with inscriptions. Finally, having scrupulously exhibited what they saw, the designers let themselves go and in the last plate of each series restored the entire structure in the mind's eye.

These creations were the works not of artists or archaeologists but of engineers and a few architects. They were very young engineers, recent graduates and some undergraduates of the École Polytechnique, founded in 1794, where drafting and surveying were major subjects. Equipped with drawing board, graph paper, pencil, ruler and compass, trained engineers were able to produce a sketch of any structure. The drawing could be developed into a finished picture after they had measured all dimensions. The completed engravings re-create the experience of standing before the facade of Karnak or gazing across the sands at the Pyramids with an immediacy not felt in perusing the most modern, the most elaborate of photographic albums.

It was not, of course, for artistic purposes that Bonaparte included these well-educated youngsters in the expedition. Their main responsibility was to build or mend fortifications, roads, bridges, canals and public works. Indeed, the men did discharge their mundane tasks, as they would have done

RUINS OF THE SOUTH GATE of Karnak were depicted as they stood when the engineers came upon them. The remains of the small temples of Apet and Khons can be seen in the foreground, and part of the great temple lies in the center background. The first plate of each architectural series in *La Description de l'Égypte* was devoted to an accurate illustration of a monument as it appeared in 1799.

in France. Nevertheless, Egypt was the great adventure of their lives. One team even succeeded in the archaeological feat of excavating the route of the canal that had linked the Red Sea to the Mediterranean in ancient times. (They had the misfortune, however, to calculate that sea level at the former end was 33 feet higher than at the latter—a conclusion that was dead wrong because sea level is sea level the world over.)

The men of the expedition had known nothing of the country when they embarked from France, not even that Egypt was their destination. That information had been kept secret from all but the high command. The members of the Commission of Science and Arts had no guides to the symbolism and significance of what they saw other than the historians and geographers of antiquity: Herodotus, Strabo and Diodorus of Sicily. The elementary facts provided to tourists in the most superficial of modern guidebooks were unknown to them. They supposed small structures to be shrines, middling ones to be temples and the greatest to be palaces. They took the crowns of Upper and Lower Egypt for elaborate coiffures.

Even so, confronted with hundreds of bas-reliefs and thousands of hieroglyphs, the young engineers copied the lot so faithfully that in many instances they preserved the evidence of inscriptions and structures that have since disappeared. For example, the temple of Isis across the Nile from Esna was destroyed in 1828 during the regime of

SOUTH GATE OF KARNAK was drawn by members of Napoleon's Commission of Science and Arts as they imagined it to have been originally. The scene might have served as a stage set for the grand march in Verdi's *Aida:* while the populace looks on, a Theban king passes through the triumphal arch, preceded by retainers and followed by his prisoners.

TEMPLE OF ISIS, across the Nile River from the ruins of Esna, was destroyed in 1828 during the regime of Mehemet Ali, the modernizer of Egypt. *La Description de l'Égypte* is an important resource for archaeology because it contains several plates preserving the record of structures and inscriptions that no longer exist.

Mehemet Ali. It could be said that Egyptology began with *La Description de l'Égypte*—except that the authors had no clue to the meaning of what they were recording. The opportunity afforded the admirer of their work is, therefore, unique in the history of science. These plates display the subject matter of a science in the absence of the science. It was only in 1822 that Jean-François Champollion succeeded in matching the name Ptolemy in the three scripts—hieroglyphic, demotic and Greek—inscribed on the Rosetta stone. Not until the 1850s were scholars able to construe whole texts.

As for science in the ordinary sense, the Egyptian environment created exceptional opportunities. Monge's explanation of mirages is the most famous memoir contributed to the institute. The sight of island villages shimmering in the waters of an ever receding lake had tormented the army during the grueling march from Alexandria. In a paper read before his colleagues on August 28, 1798, four weeks after Cairo was taken, Monge interpreted the illusion as the effect of light rays from beyond the horizon reflected from the surface of a layer of air superheated at ground level by the sun-soaked sand. Although modern optics attributes the effect to a dual refraction within the surface layer, Monge had the underlying physics right.

A memoir by his fellow senior scientist, Berthollet, had greater consequence, both for the author's career and for his science. Berthollet's *Observations sur le natron* may be considered the point of departure for physical chemistry. Highly saline lakes in a dried-up river basin some 60 miles west of Cairo were known then by the name "natron," Greek for "soda." Surrounding them are limestone formations on which deposits of that commodity occurred naturally. Interspersed among those patches were stretches where clay predominated—in those areas the soil was full of salt and free of soda. Berthollet deduced that in the limestone sectors, the lime (calcium carbonate) decomposed salt (sodium chloride) in the presence of heat and humidity. The resulting encrustation of natron (sodium carbonate) dried out and solidified on the surface. The accompanying product, calcium chloride, being extremely deliquescent, took up water and seeped away into the ground.

The significant feature of Berthollet's finding was that the reaction known in the laboratory was the exact reverse. Chemists concerned with affinities between substances normally supposed that the chemical nature of reagents was what controlled the direction of a reaction. Here, though, was an instance in which physical factors predominated. Berthollet began the paper reporting these observations in Egypt and completed it in Paris. There he developed the argument into the central theme of his major work published in 1803, *Essai de statique chimique,* which treats the effects of pressure, heat, light and the relative concentration of reagents in determining the course of reactions.

It was, however, the young naturalists, rather than the two senior scientists, whose presence in Egypt might have been expected to make a significant difference to their discipline. Twelve in number, they made up the second largest contingent of the expedition, after the engineers, and were investigating a flora and fauna unknown in Europe. Indeed, two of the authors did make names for themselves. When they embarked for Egypt, Étienne Geoffroy Saint-Hilaire was in the early stage of his career, and Jules-César Lelorgne de Savigny was at the very beginning of his. Savigny took responsibility for invertebrate zoology and for ornithology as well as for a few reptiles; Geoffroy covered all the other vertebrates.

Geoffroy and Savigny were naturalists of similar interests and very dissimilar scientific personalities. In contrast to Georges Cuvier, not much their elder but already dominant in the Paris Museum of Natural History, both were zoologists whose research moved beyond taxonomy, the work of classification, to morphology, the study of form and structure. The former had been the main preoccupation of the natural history of the 18th century; the latter became an important subdiscipline of the emerging science of biology in the 19th century. Geoffroy made the transition in the spirit of romanticism and Savigny in service to precision.

Geoffroy was of a generous, even effusive disposition. His letters to his colleagues of the museum, and especially to Cuvier (who had refused to join the expedition), are almost embarrassing in their protestations of friendship—the more so because his pleas for assurance that he had not been forgotten went unanswered. Geoffroy also had an eye for novelty. The more spectacular the creature, the more eagerly he described and dissected it. The crocodile, the great Nile tortoise, the *polyptère bichir* (a lungfish with 16 dorsal fins), the torpedo ray and the thunder-fish were among the dramatic forms he opened with his scalpel. In one respect, Geoffroy's style is reminiscent of the great Georges-Louis Leclerc de Buffon's in the

preceding century. Geoffroy's accounts include character sketches of the animals—their habits, their conduct, almost their morality. His anatomies were, however, highly skilled. The detail is exact. The drawings and descriptions are clear. He knew the literature.

The morphological direction that Geoffroy's interests were taking became evident in three memoirs on the anatomy of fish published in 1807. He had just had a revelation, Geoffroy wrote, while working on his ichthyology for *La Description de l'Égypte.* Until then, he had concurred with the accepted opinion among naturalists that in important respects the internal organization of fish was categorically different from that of vertebrates generally. Now, on closely examining his Egyptian specimens and Cuvier's collection, he reports that he is thrilled to find that the very organs that had most stubbornly resisted comparison actually exhibit profound analogies with the parts of other vertebrates.

The shift toward morphology led Geoffroy to compose his principal work, *Philosophie anatomique,* between 1818 and 1822. His argument is that differences in the organization of all classes of vertebrates represent variations on a fundamental unity of plan, a notion he later extended to invertebrates. The extravagance of these ideas was in conflict with Cuvier's commitment to the fixity of species and involved the two

CIRCULAR ZODIAC OF DENDARA had been set into the ceiling of a shrine off the Osiris Chapel of the temple. The two engineers who produced the drawing had to work by candlelight lying flat on their backs in the gloom of a closed chamber. In 1821 this masterpiece was moved to Paris, where it is now on display in the Louvre.

former friends in a notorious confrontation in 1830.

Unlike Geoffroy, Savigny first made his name through a small work of broad interest, and only much later did he move from generalization toward specialization. His *Histoire naturelle et mythologique de l'ibis,* published in 1805, is a charming combination of classical erudition and zoological precision. The veneration of the white ibis in ancient Egypt reflected its supposed appetite for flying snakes, which, according to legend, would have otherwise invaded the land. In fact, Egypt was in no danger from snakes, winged or earthbound, except as symbols of evil. Moreover, the ibis is a wading bird and eats no snakes. The real source of its sacred character was its arrival on the summer winds. It reappeared annually as the harbinger of life-giving waters, hence its identification with Toth, the ibis-headed equivalent of Mercury. Savigny notes that if the stomach cavities of ibis mummies held the remains of snakes, and typically they did, it was because the embalmers had served truths deeper than mere facts of natural history.

His book a fine success, Savigny settled down to put his Egyptian specimens in order. Still, he found himself at a loss to ascribe distinguishing characters to the manifold types of insects and crustaceans he had collected. No entomologist had yet identified systems of organs generally disposed in a regular manner—as Linnaeus had done with the sex organs of plants—so that variations might be compared from species to species and genus to genus. Working with some 1,500 specimens, Savigny began the search by detaching the external features and making separate drawings of each. Few of the creatures were as much as a centimeter long, and most of them were much smaller. A survey of his thousands of drawings yielded the key to classification. Because the same elements of mouthparts occurred in all forms, the modifications of these structures afforded the most reliable comparisons between species.

NATURAL HISTORY PLATES and monographs depict the flora, fauna and mineral species of the Nile valley. Étienne Geoffroy Saint-Hilaire, who drew this lungfish (*top*), had a penchant for creatures of unusual or extravagant form. Drawings by the mining engineer François-Michel de Rozière, such as this sample of breccia (*right*), set a new standard for precise geological illustration.

Savigny devoted his first paper to moths and butterflies, the most controversial case. In this report he took issue with his seniors, Cuvier and the foremost entomologist in France, Pierre André Latreille, both of whom considered that the jaws of the caterpillar disappear on its metamorphosis into a butterfly. Not so, Savigny found. He was able to discern forms of miniature lips, mandibles and jaws so modified as to be virtually unrecognizable—a finding for which Cuvier and Latreille gave him full credit. These criteria enabled Savigny to establish the morphological definition of the class of insects proper: the hexapods, which have six legs and two antennae.

In his next memoir, Savigny turned to the second great division of articulated invertebrates—the myriapods (including centipedes), arachnids and crustaceans—which Linnaeus had lumped together under the designation "insect." Mouthparts are again the key to classification. So extraordinary were the variations that Savigny adduced homologies with a daring and virtuosity quite uncharacteristic of the staid world of taxonomy. In certain groups, such as crabs, organs that serve for mastication are comparable to those that other orders use for locomotion: what are feet in hexapods appear to be transformed into jaws in crabs. Later work on the ascidians, or tunicate worms, was no less startling. Savigny showed that creatures vaguely called zoophytes, far from being instances of extreme simplicity, exhibit complex colonial arrangements. A final study on annelids advanced the systematization of Cuvier's class of red-blooded worms.

Savigny's work, in short, marks the beginning of zoological studies of homology in general. At the same time, his accuracy in detail was such that his plates on mollusks were reprinted as late as 1926, not for antiquarian but for scientific reasons. He never completed a comprehensive treatise, however; Savigny was unable even to prepare the annotations to accompany his plates in *La Description de l'Égypte.* As he worked on their infinitely fine detail, he suffered recurrent attacks of a neurological disorder that robbed him of effective eyesight when it set in permanently in 1824. He attributed the problem to the late onset of the ophthalmia that had afflicted many members of the Egyptian expedition. In fact, in the diagnosis of modern specialists, a temporal mode epilepsy may probably have been responsible. Unable to support the light of day, Savigny passed the last 30 years of his life enveloped in a black veil whenever the shutters were opened. His only remaining publication was a taxonomy of the highly systematic hallucinations produced by the turbulence in his optic nerve, an aurora borealis inside his head.

The botany in *La Description de l'Égypte* is a little disappointing when compared with the zoology, but the mineralogy is intriguing indeed. There are 15 magnificent plates, comprising more than 100 illustrations of the petrology of Egypt, together with an extensive monograph on the physical geography of the country. The author, François-Michel de Rozière, was a mining engineer who made no other contribution to formal science and scholarship. Like Savigny, he treated highly

specialized subject matter in a manner that held general interest.

Mineralogy was even then differentiating itself from natural history and entering into the emerging discipline of geology. Rozière designed his plates expressly to exemplify the importance of the graphic arts for the new science. Geologists had yet to develop a language like that of chemistry or botany that permitted identifying minerals by specific names. The descriptions of rocks in geologic writings were meaningless in the absence of the specimen in question. A properly executed illustration could supply the lack, and Rozière took pains to ensure that his drawings were not merely pictures of the particular rocks on his table at the moment but schematic renderings of all the distinguishing features of the type each one represented. The elements were to be written down in a description, but the form, the color, the mixture and, above all, the texture—in short, the properties required for recognition—needed to be shown graphically.

Rozière's monograph is entitled *On the Physical Constitution of Egypt and on Its Relation with the Ancient Institutions of the Country.* Had the author been a philosopher or an ideologist of some sort, the argument would have been thought daring because his purpose was to show how culture derives from material circumstances and not from divine dispensation or other transcendental factors. Composed by this mining engineer, the treatment is merely matter of fact. In no other country, Rozière observes, has a highly developed society such as that of ancient Egypt ever exhibited such dependence on a single set of physical factors. Everything in the laws of the land and the customs of the people derives from the behavior of the Nile. The rise and fall of the river not only shaped the civilization of Egypt but also accounted for the influence of its culture on the theogonies, the sciences, and the arts and crafts of all antiquity. The phenomenon, moreover, is one that can be studied in an isolation comparable to that of a laboratory, which for these purposes Egypt was.

It is in part two of *La Description de l'Égypte,* concerned with the country as it was at the end of the 18th century, that the work most largely fulfills the promise of its title. Memoirs and stud-

DIVINE HARPIST from the wall of the tomb of Ramses the Third is among the 50 color illustrations in the document.

ies on topography occupy much of the text. The engineering parties that ran the traverses from which maps were constructed had a mission beyond mere cartography. Surveyors fanning out through all the villages across the delta and up the Nile were instructed to take what amounted to a census, reporting the number of inhabitants and families, their status and occupations, the mode of agriculture, the population of horses and camels, the practice of animal husbandry, the types of commerce and industry, the location of quarries, of oases, of canals, of towpaths, the means of transport and communication as well as the ethnic and religious character of the people, both settled and nomadic.

BLACK AND WHITE IBISES were illustrated by Jules-César Lelorgne de Savigny, a founder of morphology. His book on the natural history of the ibis notes that the white ibis, venerated by the Egyptians for protecting their land from serpents, never eats snakes. Ancient embalmers respected and conserved the myth, however, by placing snakes in the stomach cavities of the birds they mummified.

Topography was a subject of medical and physical concern because the goal of 18th-century medicine was to balance the environment and the physiological constitution of men, women and children. The head physician of the expedition was Nicolas Desgenettes, and the head surgeon was Dominique Jean Larrey. The Egyptian setting being dramatic, Desgenettes composed a *Topographie physique et médicale de l'Égypte,* which included collaboration from the astronomer Nicolas-Antoine Nouet. Throughout the occupation, Desgenettes assembled data on the population dynamics of Egypt, compiled a necrology of Cairo for the three years the French were in control and composed a classic of military medicine that set out policies for sanitation, public health and the organization of hospitals.

Larrey, for his part, wrote mainly about disease. He gave clinical descriptions of ophthalmia (usually trachoma), bubonic plague, tetanus, yellow fever, leprosy, elephantiasis, and testicular atrophy and gigantism. In his view, the etiology of plague, fever and tetanus clearly involved a specific external agent, for which he sometimes used the word "virus" and sometimes "germ." His concept of disease was as specific and objective as anything that entered into 19th-century medicine from the new clinical practice in Paris, an approach that he appears to have anticipated and developed on his own.

Among the memoirs and monographs in part two of *La Description de l'Égypte* are many covering topics that nowadays would be classified as social science or humanities. These include anthropology (both cultural and physical), demography, meteorology, political science, sociology, geopolitics, agronomy, microeconomics, medieval history, administrative history, linguistics and musicology—disciplines that did not yet exist for the most part. The authors of these pieces also were engineers, scientists and military men, people trained to be systematic, who knew how to look around them and take the measure of what they saw.

Their attitude is that of observers of phenomena. They often said to one another that no other country in the world, and certainly not France, had ever been the subject of such thorough study as theirs on Egypt. That fact began to change when they returned to France. Most of them continued in the service of the state. The voracious fact-gathering they had practiced in Egypt became characteristic both of the Napoleonic regime and the monarchy restored in 1815. One engineer, Chabrol de Volvic, is a fair example. As a youngster, he designed many of the plates on antiquity and composed an essay on the customs of the modern inhabitants of Egypt. Chabrol finished his career as prefect of the Seine in the 1820s. He then ordered the compilation of an urban topography, *Statistique de la Ville de Paris,* in effect an application to the capital of France of the techniques of deep description he and his colleagues had employed in Egypt.

Above and beyond the enormous compilation of information on Egypt, the significance of the participation of science in the expedition lies in the relation it portended between formal knowledge and politics. Unlike the mercantile colonialism that preceded it, the occupation of Egypt had a cultural component. Technical competence was at the forefront of culture. Bonaparte understood that point, not abstractly but intuitively, as he understood whatever related to the exercise of power. His was the impulse that implanted a clone of French science on the banks of the Nile. The British in India, the Dutch in Indonesia, the Spaniards and Portuguese in America—earlier imperialism had not attempted anything of the kind. The spread of European science and its appurtenances to African and Asian societies under the aegis of military conquest and political power began with the French conquest of Egypt.

The motivation that Fourier ascribes to Bonaparte in the preface may be read as a prophetic rationale: "He was aware of the influence that this event [the conquest of Egypt] would have on the relations of Europe with the East and with the interior of Africa as well as on maritime affairs in the Mediterranean and the future of Asia. He set himself the goals of abolishing the tyranny of the Mamelukes, of extending irrigation and agriculture, of instituting regular commerce between the Mediterranean and the Arabian Sea, of fostering commercial enterprises, of offering useful examples of European industry to the Orient, and finally of improving the standard of living of the inhabitants and procuring them all the advantages of an improved civilization. These objectives would be unattainable without the continual application of science and the technical arts."

FURTHER READING

SCIENCE AND POLITY IN FRANCE AT THE END OF THE OLD REGIME. Charles C. Gillispie. Princeton University Press, 1980.

SCIENTIFIC ASPECTS OF THE FRENCH EGYPTIAN EXPEDITION: 1798–1801. Charles C. Gillispie in *Proceedings of the American Philosophical Society,* Vol. 133, No. 4, pages 447–474; December 1989.

MONUMENTS OF EGYPT: THE COMPLETE ARCHAEOLOGICAL PLATES FROM THE *DESCRIPTION DE L'EGYPTE.* Fourth printing. Edited by Charles C. Gillispie and Michel Dewachter. Princeton Architectural Press, 1994.

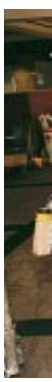
Software's Chronic Crisis

by W. Wayt Gibbs, *staff writer*

Denver's new international airport was to be the pride of the Rockies, a wonder of modern engineering. Twice the size of Manhattan, 10 times the breadth of Heathrow, the airport is big enough to land three jets simultaneously—in bad weather. Even more impressive than its girth is the airport's subterranean baggage-handling system. Tearing like intelligent coal-mine cars along 21 miles of steel track, 4,000 independent "telecars" route and deliver luggage between the counters, gates and claim areas of 20 different airlines. A central nervous system of some 100 computers networked to one another and to 5,000 electric eyes, 400 radio receivers and 56 bar-code scanners orchestrates the safe and timely arrival of every valise and ski bag.

At least that is the plan. For nine months, this Gulliver has been held captive by Lilliputians—errors in the software that controls its automated baggage system. Scheduled for takeoff by last Halloween, the airport's grand opening was postponed until December to allow BAE Automated Systems time to flush the gremlins out of its $193-million system. December yielded to March. March slipped to May. In June the airport's planners, their bond rating demoted to junk and their budget hemorrhaging red ink at the rate of $1.1 million a day in interest and operating costs, conceded that they could not predict when the baggage system would stabilize enough for the airport to open.

To veteran software developers, the Denver debacle is notable only for its visibility. Studies have shown that for every six new large-scale software systems that are put into operation, two others are canceled. The average software development project overshoots its schedule by half; larger projects generally do worse. And



***SOFTWARE GLITCHES** in an automated baggage-handling system force Denver International Airport to sit empty nine months after airplanes were to fill these gates and runways* **(top).** *The sys-*

Despite 50 years of progress, the software industry remains years—perhaps decades—short of the mature engineering discipline needed to meet the demands of an information-age society

tem that is supposed to shunt luggage in 4,000 independent "telecars" along 21 miles of track still opened, damaged and misrouted cargo as testing continued in July **(bottom).**

some three quarters of all large systems are "operating failures" that either do not function as intended or are not used at all.

The art of programming has taken 50 years of continual refinement to reach this stage. By the time it reached 25, the difficulties of building big software loomed so large that in the autumn of 1968 the NATO Science Committee convened some 50 top programmers, computer scientists and captains of industry to plot a course out of what had come to be known as the software crisis. Although the experts could not contrive a road map to guide the industry toward firmer ground, they did coin a name for that distant goal: software engineering, now defined formally as "the application of a systematic, disciplined, quantifiable approach to the development, operation and maintenance of software."

A quarter of a century later software engineering remains a term of aspiration. The vast majority of computer code is still handcrafted from raw programming languages by artisans using techniques they neither measure nor are able to repeat consistently. "It's like musket making was before Eli Whitney," says Brad J. Cox, a professor at George Mason University. "Before the industrial revolution, there was a nonspecialized approach to manufacturing goods that involved very little interchangeability and a maximum of craftsmanship. If we are ever going to lick this software crisis, we're going to have to stop this hand-to-mouth, every-programmer-builds-everything-from-the-ground-up, preindustrial approach."

The picture is not entirely bleak. Intuition is slowly yielding to analysis as programmers begin using quantitative measurements of the quality of the software they produce to improve

the way they produce it. The mathematical foundations of programming are solidifying as researchers work on ways of expressing program designs in algebraic forms that make it easier to avoid serious mistakes. Academic computer scientists are starting to address their failure to produce a solid corps of software professionals. Perhaps most important, many in the industry are turning their attention toward inventing the technology and market structures needed to support interchangeable, reusable software parts.

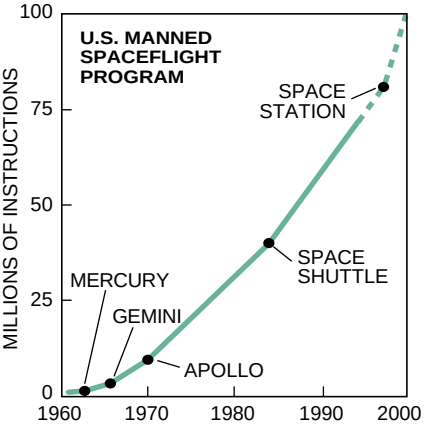
"Unfortunately, the industry does not uniformly apply that which is well-known best practice," laments Larry E. Druffel, director of Carnegie Mellon University's Software Engineering Institute. In fact, a research innovation typically requires 18 years to wend its way into the repertoire of standard programming techniques. By combining their efforts, academia, industry and government may be able to hoist software development to the level of an industrial-age engineering discipline within the decade. If they come up short, society's headlong rush into the information age will be halting and unpredictable at best.

Shifting Sands

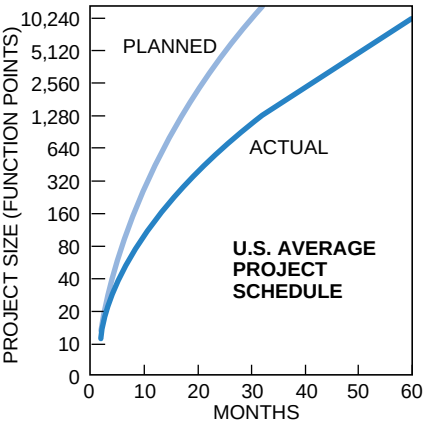
"We will see massive changes [in computer use] over the next few years, causing the initial personal computer revolution to pale into comparative insignificance," concluded 22 leaders in software development from academia, industry and research laboratories this past April. The experts gathered at Hedsor Park, a corporate retreat near London, to commemorate the NATO conference and to analyze the future directions of software. "In 1968 we knew what we wanted to build but couldn't," reflected Cliff Jones, a professor at the University of Manchester. "Today we are standing on shifting sands."

The foundations of traditional programming practices are eroding swiftly, as hardware engineers churn out ever faster, cheaper and smaller machines. Many fundamental assumptions that programmers make—for instance, their acceptance that everything they produce will have defects—must change in response. "When computers are embedded in light switches, you've got to get the software right the first time because you're not going to have a chance to update it," says Mary M. Shaw, a professor at Carnegie Mellon.

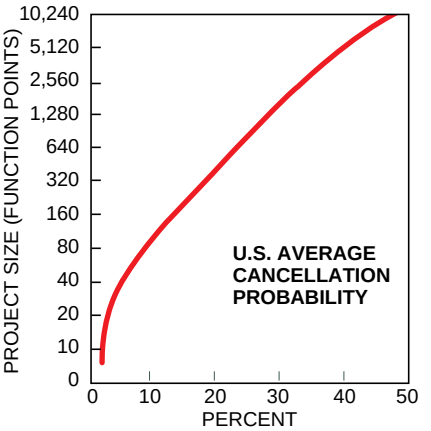
"The amount of code in most consumer products is doubling every two years," notes Remi H. Bourgonjon, director of software technology at Philips Research Laboratory in Eindhoven. Already, he reports, televisions may contain up to 500 kilobytes of software; an electric shaver, two kilobytes. The power trains in new General Motors cars run 30,000 lines of computer code.



SOURCE: Barry W. Boehm



SOURCE: Software Productivity Research



SOURCE: Software Productivity Research

SOFTWARE IS EXPLODING in size as society comes to rely on more powerful computer systems (*top*). That faith is often rewarded by disappointment as most large software projects overrun their schedules (*middle*) and many fail outright (*bottom*)—usually after most of the development money has been spent.

Getting software right the first time is hard even for those who care to try. The Department of Defense applies rigorous—and expensive—testing standards to ensure that software on which a mission depends is reliable. Those standards were used to certify *Clementine,* a satellite that the DOD and the National Aeronautics and Space Administration directed into lunar orbit this past spring. A major part of the Clementine mission was to test targeting software that could one day be used in a space-based missile defense system. But when the satellite was spun around and instructed to fix the moon in its sights, a bug in its program caused the spacecraft instead to fire its maneuvering thrusters continuously for 11 minutes. Out of fuel and spinning wildly, the satellite could not make its rendezvous with the asteroid Geographos.

Errors in real-time systems such as *Clementine* are devilishly difficult to spot because, like that suspicious sound in your car engine, they often occur only when conditions are just so [see "The Risks of Software," by Bev Littlewood and Lorenzo Strigini; SCIENTIFIC AMERICAN, November 1992]. "It is not clear that the methods that are currently used for producing safety-critical software, such as that in nuclear reactors or in cars, will evolve and scale up adequately to match our future expectations," warned Gilles Kahn, the scientific director of France's INRIA research laboratory, at the Hedsor Park meeting. "On the contrary, for real-time systems I think we are at a fracture point."

Software is buckling as well under tectonic stresses imposed by the inexorably growing demand for "distributed systems": programs that run cooperatively on many networked computers. Businesses are pouring capital into distributed information systems that they hope to wield as strategic weapons. The inconstancy of software development can turn such projects into Russian roulette.

Many companies are lured by goals that seem simple enough. Some try to reincarnate obsolete mainframe-based software in distributed form. Others want to plug their existing systems into one another or into new systems with which they can share data and a friendlier user interface. In the technical lingo, connecting programs in this way is often called systems integration. But Brian Randell, a computer scientist at the University of Newcastle upon Tyne, suggests that "there is a better word than integration, from old R.A.F. slang: namely, 'to graunch,' which means 'to make to fit by the use of excessive force.'"

It is a risky business, for although

software seems like malleable stuff, most programs are actually intricate plexuses of brittle logic through which data of only the right kind may pass. Like handmade muskets, several programs may perform similar functions and yet still be unique in design. That makes software difficult to modify and repair. It also means that attempts to graunch systems together often end badly.

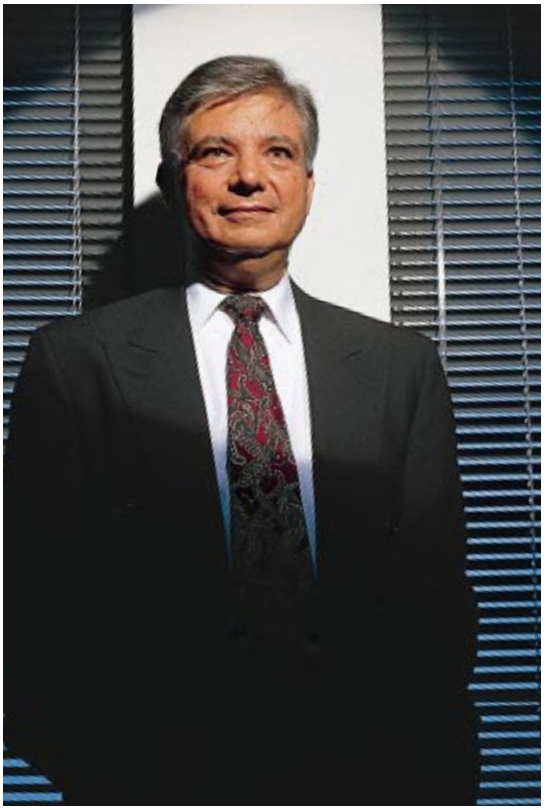
In 1987, for example, California's Department of Motor Vehicles decided to make its customers' lives easier by merging the state's driver and vehicle registration systems—a seemingly straightforward task. It had hoped to unveil convenient one-stop renewal kiosks last year. Instead the DMV saw the projected cost explode to 6.5 times the expected price and the delivery date recede to 1998. In December the agency pulled the plug and walked away from the seven-year, $44.3-million investment.

Sometimes nothing fails like success. In the 1970s American Airlines constructed SABRE, a virtuosic, $2-billion flight reservation system that became part of the travel industry's infrastructure. "SABRE was the shining example of a strategic information system because it drove American to being the world's largest airline," recalls Bill Curtis, a consultant to the Software Engineering Institute.

Intent on brandishing software as effectively in this decade, American tried to graunch its flight-booking technology with the hotel and car reservation systems of Marriott, Hilton and Budget. In 1992 the project collapsed into a heap of litigation. "It was a smashing failure," Curtis says. "American wrote off $165 million against that system."

The airline is hardly suffering alone. In June IBM's Consulting Group released the results of a survey of 24 leading companies that had developed large distributed systems. The numbers were unsettling: 55 percent of the projects cost more than expected, 68 percent overran their schedules and 88 percent had to be substantially redesigned.

The survey did not report one critical statistic: how reliably the completed programs ran. Often systems crash because they fail to expect the unexpected. Networks amplify this problem. "Distributed systems can consist of a great set of interconnected single points of failure, many of which you have not identified beforehand," Randell explains. "The complexity and fragility of these systems pose a major challenge."



EXPERIMENTALIST Victor R. Basili helped found the Software Engineering Laboratory to push programming onto a firmer foundation of mathematics and science.

The challenge of complexity is not only large but also growing. The bang that computers deliver per buck is doubling every 18 months or so. One result is "an order of magnitude growth in system size every decade—for some industries, every half decade," Curtis says. To keep up with such demand, programmers will have to change the way that they work. "You can't build skyscrapers using carpenters," Curtis quips.

Mayday, Mayday

When a system becomes so complex that no one manager can comprehend the entirety, traditional development processes break down. The Federal Aviation Administration (FAA) has faced this problem throughout its decade-old attempt to replace the nation's increasingly obsolete air-traffic control system [see "Aging Airways," by Gary Stix; SCIENTIFIC AMERICAN, May].

The replacement, called the Advanced Automation System (AAS), combines all the challenges of computing in the 1990s. A program that is more than a million lines in size is distributed across hundreds of computers and embedded into new and sophisticated hardware, all of which must respond around the clock to unpredictable real-time events. Even a small glitch potentially threatens public safety.

To realize its technological dream, the FAA chose IBM's Federal Systems Company, a well-respected leader in software development that has since been purchased by Loral. FAA managers expected (but did not demand) that IBM would use state-of-the-art techniques to estimate the cost and length of the project. They assumed that IBM would screen the requirements and design drawn up for the system in order to catch mistakes early, when they can be fixed in hours rather than days. And the FAA conservatively expected to pay about $500 per line of computer code, five times the industry average for well-managed development processes.

According to a report on the AAS project released in May by the Center for Naval Analysis, IBM's "cost estimation and development process tracking used inappropriate data, were performed inconsistently and were routinely ignored" by project managers. As a result, the FAA has been paying $700 to $900 per line for the AAS software. One reason for the exorbitant price is that "on average every line of code developed needs to be rewritten once," bemoaned an internal FAA report.

Alarmed by skyrocketing costs and tests that showed the half-completed system to be unreliable, FAA administrator David R. Hinson decided in June to cancel two of the four major parts of the AAS and to scale back a third. The $144 million spent on these failed programs is but a drop next to the $1.4 billion invested in the fourth and central piece: new workstation software for air-traffic controllers.

That project is also spiraling down the drain. Now running about five years late and more than $1 billion over budget, the bug-infested program is being scoured by software experts at Carnegie Mellon and the Massachusetts Institute of Technology to determine whether it can be salvaged or must be canceled outright. The reviewers are scheduled to make their report in September.

Disaster will become an increasingly common and disruptive part of software development unless programming takes on more of the characteristics of an engineering discipline rooted firmly in science and mathematics [*see box on page 92*]. Fortunately, that trend has already begun. Over the past decade in-

ALL OF FRANCE'S 6,000 electric trains will use speed- and switching-control software developed by GEC Alsthom using mathematical methods to prove that the programs are written correctly.

dustry leaders have made significant progress toward understanding how to measure, consistently and quantitatively, the chaos of their development processes, the density of errors in their products and the stagnation of their programmers' productivity. Researchers are already taking the next step: finding practical, repeatable solutions to these problems.

Proceeds of Process

In 1991, for example, the Software Engineering Institute, a software think tank funded by the military, unveiled its Capability Maturity Model (CMM). "It provides a vision of software engineering and management excellence," beams David Zubrow, who leads a project on empirical methods at the institute. The CMM has at last persuaded many programmers to concentrate on measuring the process by which they produce software, a prerequisite for any industrial engineering discipline.

Using interviews, questionnaires and the CMM as a benchmark, evaluators can grade the ability of a programming team to create predictably software that meets its customers' needs. The CMM uses a five-level scale, ranging from chaos at level 1 to the paragon of good management at level 5. To date, 261 organizations have been rated.

"The vast majority—about 75 percent—are still stuck in level 1," Curtis reports. "They have no formal process, no measurements of what they do and no way of knowing when they are on the wrong track or off the track altogether." (The Center for Naval Analysis concluded that the AAS project at IBM Federal Systems "appears to be at a low 1 rating.") The remaining 24 percent of projects are at levels 2 or 3.

Only two elite groups have earned the highest CMM rating, a level 5. Motorola's Indian programming team in Bangalore holds one title. Loral's (formerly IBM's) on-board space shuttle software project claims the other. The Loral team has learned to control bugs so well that it can reliably predict how many will be found in each new version of the software. That is a remarkable feat, considering that 90 percent of American programmers do not even keep count of the mistakes they find, according to Capers Jones, chairman of Software Productivity Research. Of those who do, he says, few catch more than a third of the defects that are there.

Tom Peterson, head of Loral's shuttle software project, attributes its success to "a culture that tries to fix not just the bug but also the flaw in the testing process that allowed it to slip through." Yet some bugs inevitably escape detection. The first launch of the space shuttle in 1981 was aborted and delayed for two days because a glitch prevented the five on-board computers from synchronizing properly. Another flaw, this one in the shuttle's rendezvous program, jeopardized the *Intelsat-6* satellite rescue mission in 1992.

Although the CMM is no panacea, its promotion by the Software Engineering Institute has persuaded a number of leading software companies that quantitative quality control can pay off in the long run. Raytheon's equipment division, for example, formed a "software engineering initiative" in 1988 after flunking the CMM test. The division began pouring $1 million per year into refining rigorous inspection and testing guidelines and training its 400 programmers to follow them.

Within three years the division had jumped two levels. By this past June, most projects—including complex radar and air-traffic control systems—were finishing ahead of schedule and under budget. Productivity has more than doubled. An analysis of avoided rework costs revealed a savings of $7.80 for every dollar invested in the initiative. Impressed by such successes, the U.S. Air Force has mandated that all its software developers must reach level 3 of the CMM by 1998. NASA is reportedly considering a similar policy.

Mathematical Re-creations

Even the best-laid designs can go awry, and errors will creep in so long as humans create programs. Bugs squashed early rarely threaten a project's deadline and budget, however. Devastating mistakes are nearly always those in the initial design that slip undetected into the final product.

Mass-market software producers, because they have no single customer to please, can take a belated and brute-force approach to bug removal: they release the faulty product as a "beta" version and let hordes of users dig up the glitches. According to Charles Simonyi, a chief architect at Microsoft, the new version of the Windows operating system will be beta-tested by 20,000 volunteers. That is remarkably effective, but also expensive, inefficient and—since mass-produced PC products make up less than 10 percent of the $92.8-billion software market in the U.S.—usually impractical.

Researchers are thus formulating several strategies to attack bugs early or to avoid introducing them at all. One idea is to recognize that the problem a system is supposed to solve always changes as the system is being built. Denver's airport planners saddled BAE with $20 million worth of changes to the design of its baggage system long after construction had begun. IBM has been similarly bedeviled by the indecision of FAA managers. Both companies naively assumed that once their design was approved, they would be left in peace to build it.

Some developers are at last shedding that illusion and rethinking software as something to be grown rather than built. As a first step, programmers are increasingly stitching together quick prototypes out of standard graphic interface components. Like an architect's

scale model, a system prototype can help clear up misunderstandings between customer and developer before a logical foundation is poured.

Because they mimic only the outward behavior of systems, prototypes are of little help in spotting logical inconsistencies in a system's design. "The vast majority of errors in large-scale software are errors of omission," notes Laszlo A. Belady, director of Mitsubishi Electric Research Laboratory. And models do not make it any easier to detect bugs once a design is committed to code.

When it absolutely, positively has to be right, says Martyn Thomas, chairman of Praxis, a British software company, engineers rely on mathematical analysis to predict how their designs will behave in the real world. Unfortunately, the mathematics that describes physical systems does not apply within the synthetic binary universe of a computer program; discrete mathematics, a far less mature field, governs here. But using the still limited tools of set theory and predicate calculus, computer scientists have contrived ways to translate specifications and programs into the language of mathematics, where they can be analyzed with theoretical tools called formal methods.

Praxis recently used formal methods on an air-traffic control project for Britain's Civil Aviation Authority. Although Praxis's program was much smaller than the FAA's, the two shared a similar design problem: the need to keep redundant systems synchronized so that if one fails, another can instantly take over. "The difficult part was guaranteeing that messages are delivered in the proper order over twin networks," recalls Anthony Hall, a principal consultant to Praxis. "So here we tried to carry out proofs of our design, and they failed, because the design was wrong. The benefit of finding errors at that early stage is enormous," he adds. The system was finished on time and put into operation last October.

Praxis used formal notations on only the most critical parts of its software, but other software firms have employed mathematical rigor throughout the entire development of a system. GEC Alsthom in Paris is using a formal method called "B" as it spends $350 million to upgrade the switching- and speed-control software that guides the 6,000 electric trains in France's national railway system. By increasing the speed of the trains and reducing the distance between them, the system can save the railway company billions of dollars that might otherwise need to be spent on new lines.

Safety was an obvious concern. So GEC developers wrote the entire design and final program in formal notation and then used mathematics to prove them consistent. "Functional tests are still necessary, however, for two reasons," says Fernando Mejia, manager of the formal development section at GEC. First, programmers do occasionally make mistakes in proofs. Secondly, formal methods can guarantee only that software meets its specification, not that it can handle the surprises of the real world.

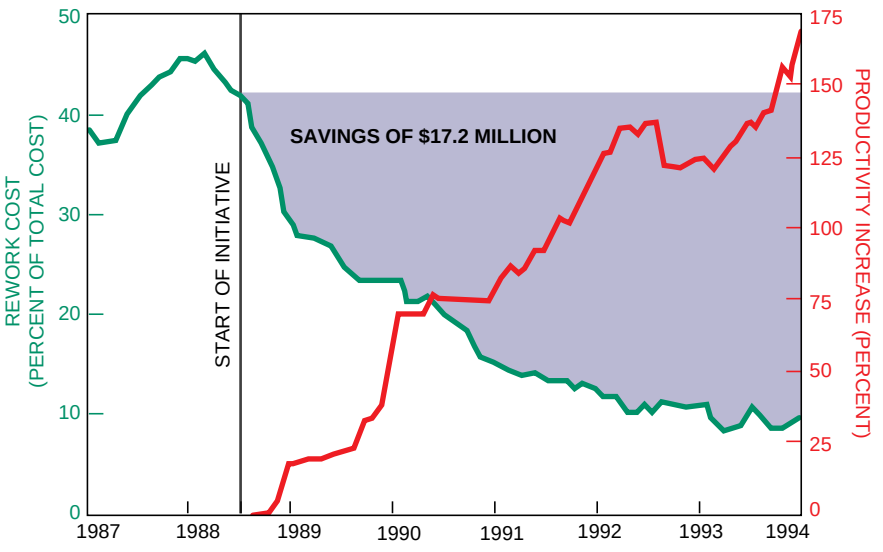
Formal methods have other problems as well. Ted Ralston, director of strategic planning for Odyssey Research Associates in Ithaca, N.Y., points out that reading pages of algebraic formulas is even more stultifying than reviewing computer code. Odyssey is just one of several companies that are trying to automate formal methods to make them less onerous to programmers. GEC is collaborating with Digilog in France to commercialize programming tools for the B method. The beta version is being tested by seven companies and institutions, including Aerospatiale, as well as France's atomic energy authority and its defense department.

On the other side of the Atlantic, formal methods by themselves have yet to catch on. "I am skeptical that Americans are sufficiently disciplined to apply formal methods in any broad fashion," says David A. Fisher of the National Institute of Standards and Technology (NIST). There are exceptions, however, most notably among the growing circle of companies experimenting with the "clean-room approach" to programming.

The clean-room process attempts to meld formal notations, correctness proofs and statistical quality control with an evolutionary approach to software development. Like the microchip manufacturing technique from which it takes its name, clean-room development tries to use rigorous engineering techniques to consistently fabricate products that run perfectly the first time. Programmers grow systems one function at a time and certify the quality of each unit before integrating it into the architecture.

Growing software requires a whole new approach to testing. Traditionally, developers test a program by running it the way they intend it to be used, which often bears scant resemblance to real-world conditions. In a clean-room process, programmers try to assign a probability to every execution path—correct and incorrect—that users can take. They then derive test cases from those statistical data, so that the most common paths are tested more thoroughly. Next the program runs through each test case and times how long it takes to fail. Those times are then fed back, in true engineering fashion, to a model that calculates how reliable the program is.

Early adopters report encouraging results. Ericsson Telecom, the European telecommunications giant, used clean-room processes on a 70-programmer project to fabricate an operating system for its telephone-switching computers. Errors were reportedly reduced to just one per 1,000 lines of program code; the industry average is about 25 times higher. Perhaps more important,

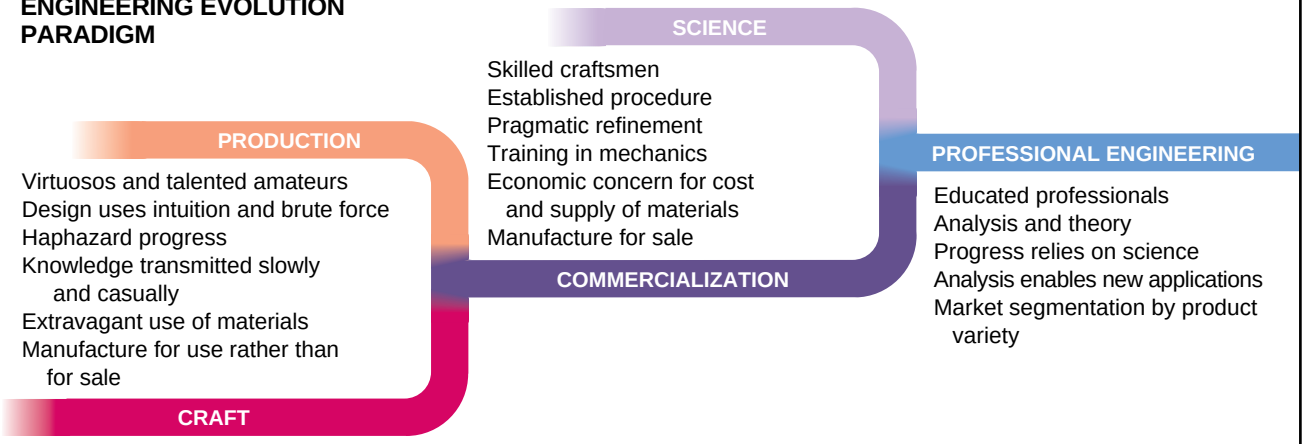


SOURCE: Raytheon

RAYTHEON HAS SAVED $17.2 million in software costs since 1988, when its equipment division began using rigorous development processes that doubled its programmers' productivity and helped them to avoid making expensive mistakes.

Progress toward Professionalism

ENGINEERING EVOLUTION PARADIGM

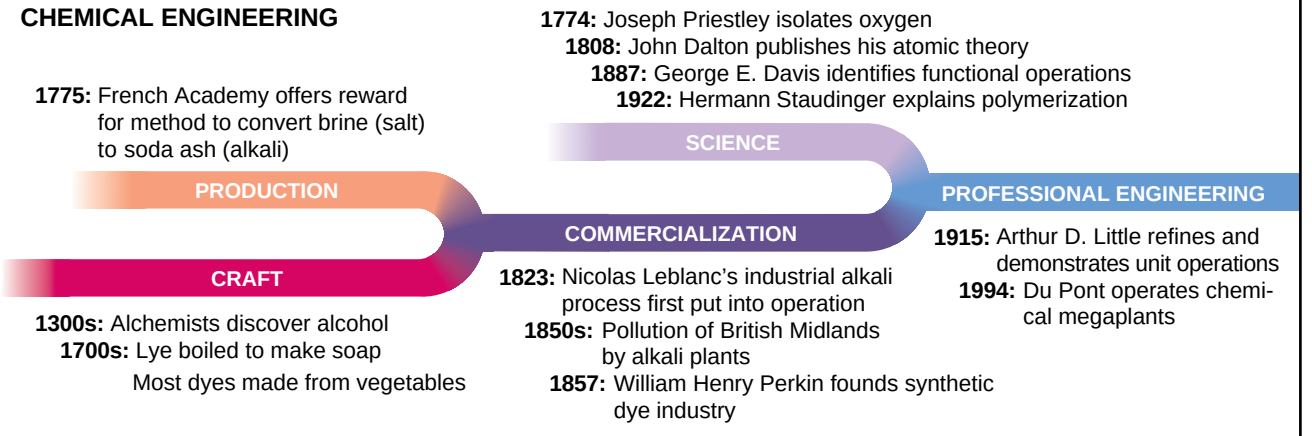


Engineering disciplines share common stages in their evolution, observes Mary M. Shaw of Carnegie Mellon University. She spies interesting parallels between software engineering and chemical engineering, two fields that aspire to exploit on an industrial scale the processes that are discovered by small-scale research.

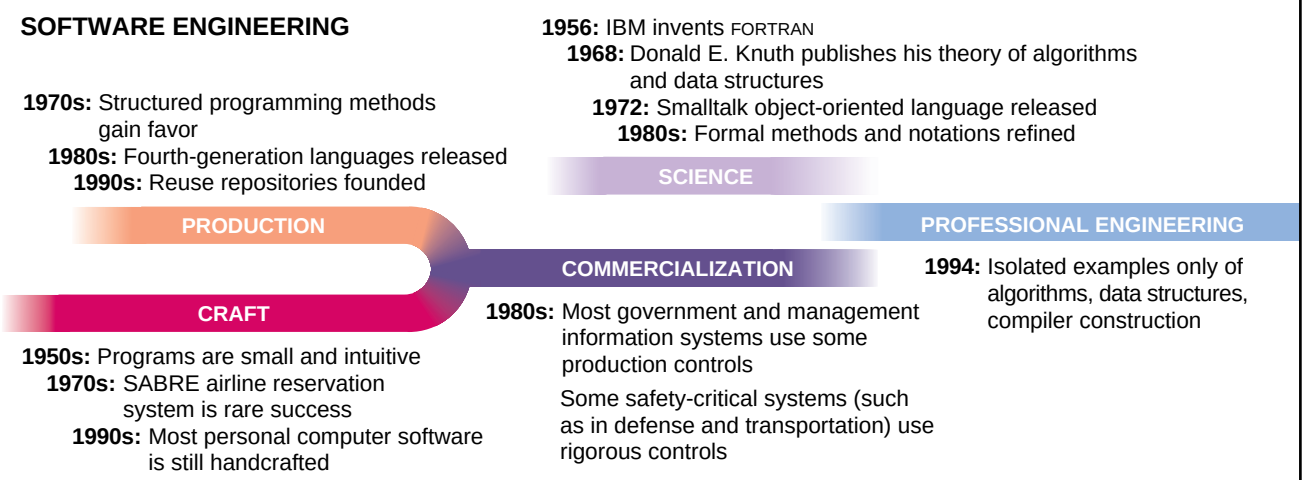
Like software developers, chemical engineers try to design processes to create safe, pure products as cheaply and quickly as possible. Unlike most programmers, however, chemical engineers rely heavily on scientific theory, mathematical modeling, proven design solutions and rigorous quality-control methods—and their efforts usually succeed.

Software, Shaw points out, is somewhat less mature, more like a cottage industry than a professional engineering discipline. Although the demand for more sophisticated and reliable software has boosted some large-scale programming to the commercial stage, computer science (which is younger than many of its researchers) has yet to build the experimental foundation on which software engineering must rest.

CHEMICAL ENGINEERING



SOFTWARE ENGINEERING



the company found that development productivity increased by 70 percent, and testing productivity doubled.

No Silver Bullet

Then again, the industry has heard tell many times before of "silver bullets" supposedly able to slay werewolf projects. Since the 1960s developers have peddled dozens of technological innovations intended to boost productivity—many have even presented demonstration projects to "prove" the verity of their boasts. Advocates of object-oriented analysis and programming, a buzzword du jour, claim their approach represents a paradigm shift that will deliver "a 14-to-1 improvement in productivity," along with higher quality and easier maintenance, all at reduced cost.

There are reasons to be skeptical. "In the 1970s structured programming was also touted as a paradigm shift," Curtis recalls. "So was CASE [computer-assisted software engineering]. So were third-, fourth- and fifth-generation languages. We've heard great promises for technology, many of which weren't delivered."

Meanwhile productivity in software development has lagged behind that of more mature disciplines, most notably computer hardware engineering. "I think of software as a cargo cult," Cox says. "Our main accomplishments were imported from this foreign culture of hardware engineering—faster machines and more memory." Fisher tends to agree: adjusted for inflation, "the value added per worker in the industry has been at $40,000 for two decades," he asserts. "We're not seeing any increases."

"I don't believe that," replies Richard A. DeMillo, a professor at Purdue University and head of the Software Engineering Research Consortium. "There has been improvement, but everyone uses different definitions of productivity." A recent study published by Capers Jones—but based on necessarily dubious historical data—states that U.S. programmers churn out twice as much code today as they did in 1970.

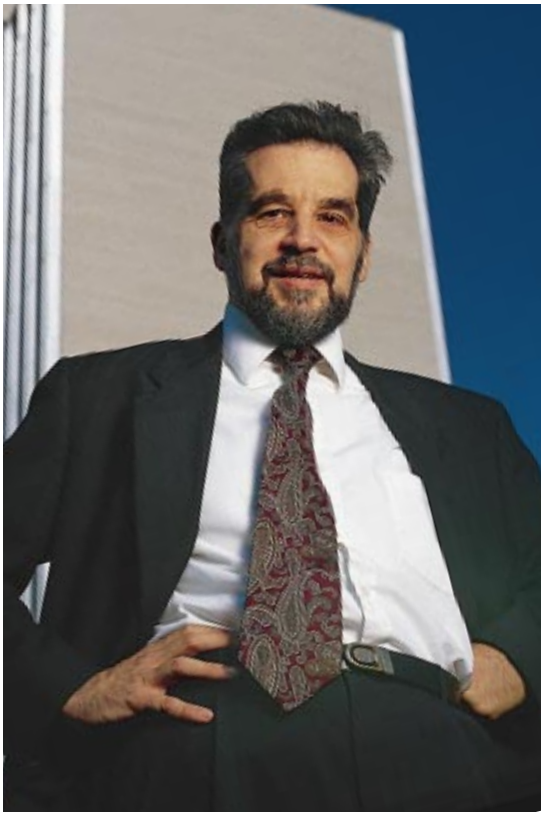
The fact of the matter is that no one really knows how productive software developers are, for three reasons. First, less than 10 percent of American companies consistently measure the productivity of their programmers.

Second, the industry has yet to settle on a useful standard unit of measurement. Most reports, including those published in peer-reviewed computer science journals, express productivity in terms of lines of code per worker per month. But programs are written in a wide variety of languages and vary enormously in the complexity of their operation. Comparing the number of lines written by a Japanese programmer using C with the number produced by an American using Ada is thus like comparing their salaries without converting from yen to dollars.

Third, Fisher says, "you can walk into a typical company and find two guys sharing an office, getting the same salary and having essentially the same credentials and yet find a factor of 100 difference in the number of instructions per day that they produce." Such enormous individual differences tend to swamp the much smaller effects of technology or process improvements.

After 25 years of disappointment with apparent innovations that turned out to be irreproducible or unscalable, many researchers concede that computer science needs an experimental branch to separate the general results from the accidental. "There has always been this assumption that if I give you a method, it is right just because I told you so," complains Victor R. Basili, a professor at the University of Maryland. "People are developing all kinds of things, and it's really quite frightening how bad some of them are," he says.

Mary Shaw of Carnegie Mellon points out that mature engineering fields codify proved solutions in handbooks so that even novices can consistently handle routine designs, freeing more talented practitioners for advanced projects. No such handbook yet exists for software, so mistakes are repeated on project after project, year after year.



AS CEO of Incremental Systems, David A. Fisher learned firsthand why software components do not sell. Now he supervises a $150-million federal program to create a market for software parts.

DeMillo suggests that the government should take a more active role. "The National Science Foundation should be interested in funding research aimed at verifying experimental results that have been claimed by other people," he says. "Currently, if it's not groundbreaking, first-time-ever-done research, program officers at the NSF tend to discount the work." DeMillo knows whereof he speaks. From 1989 to 1991 he directed the NSF's computer and computation research division.

Yet "if software engineering is to be an experimental science, that means it needs laboratory science. Where the heck are the laboratories?" Basili asks. Because attempts to scale promising technologies to industrial proportions so often fail, small laboratories are of limited utility. "We need to have places where we can gather data and try things out," DeMillo says. "The only way to do that is to have a real software development organization as a partner."

There have been only a few such partnerships. Perhaps the most successful is the Software Engineering Laboratory, a consortium of NASA's Goddard Space Flight Center, Computer Sciences Corp. and the University of Maryland. Basili helped to found the laboratory in 1976. Since then, graduate students and NASA programmers have collaborated on "well over 100 projects," Basili says, most having to do with building ground-support software for satellites.

Just Add Water

Musket makers did not get more productive until Eli Whitney figured out how to manufacture interchangeable parts that could be assembled by any skilled workman. In like manner, software parts can, if properly standardized, be reused at many different scales. Programmers have for decades used li-

A Developing World

Since the invention of computers, Americans have dominated the software market. Microsoft alone produces more computer code each year than do any of 100 nations, according to Capers Jones of Software Productivity Research in Burlington, Mass. U.S. suppliers hold about 70 percent of the worldwide software market.

But as international networks sprout and large corporations deflate, India, Hungary, Russia, the Philippines and other poorer nations are discovering in software a lucrative industry that requires the one resource in which they are rich: an underemployed, well-educated labor force. American and European giants are now competing with upstart Asian development companies for contracts, and in response many are forming subsidiaries overseas. Indeed, some managers in the trade predict that software development will gradually split between Western software engineers who design systems and Eastern programmers who build them.

"In fact, it is going on already," says Laszlo A. Belady, director of Mitsubishi Electric Research Laboratory. AT&T, Hewlett-Packard, IBM, British Telecom and Texas Instruments have all set up programming teams in India. The Pact Group in Lyons, France, reportedly maintains a "software factory" in Manila. "Cadence, the U.S. supplier of VLSI design tools, has had its software development sited on the Pacific rim for several years," reports Martyn Thomas, chairman of Praxis. "ACT, a U.K.-based systems house, is using Russian programmers from the former Soviet space program," he adds.

So far India's star has risen fastest. "Offshore development [work commissioned in India by foreign companies] has begun to take off in the past 18 to 24 months," says Rajendra S. Pawar, head of New Delhi–based NIIT, which has graduated 200,000 Indians from its programming courses (*photograph*). Indeed, India's software exports have seen a compound annual growth of 38 percent over the past five years; last year they jumped 60 percent—four times the average growth rate worldwide.

About 58 percent of the $360-million worth of software that flowed out of India last year ended up in the U.S. That tiny drop hardly makes a splash in a $92.8-billion market. But several trends may propel exports beyond the $1-billion mark as early as 1997.

The single most important factor, Pawar asserts, is the support of the Indian government, which has eased tariffs and restrictions, subsidized numerous software technology parks and export zones, and doled out five-year tax exemptions to software exporters. "The opening of the Indian economy is acting as a very big catalyst," Pawar says.

It certainly seems to have attracted the attention of large multinational firms eager to reduce both the cost of the software they need and the amount they build in-house. The primary cost of software is labor. Indian programmers come so cheap—$125 per unit of software versus $925 for an American developer, according to Jones—that some companies fly an entire team to the U.S. to work on a project. More than half of India's software exports come from such "body shopping," although tightened U.S. visa restrictions are stanching this flow.

Another factor, Pawar observes, is a growing trust in the quality of overseas project management. "In the past two years, American companies have become far more comfortable with the offshore concept," he says. This is a result in part of success stories from leaders like Citicorp, which develops banking systems in Bombay, and Motorola, which has a top-rated team of more than 150 programmers in

braries of subroutines to avoid rewriting the same code over and over. But these components break down when they are moved to a different programming language, computer platform or operating environment. "The tragedy is that as hardware becomes obsolete, an excellent expression of a sorting algorithm written in the 1960s has to be rewritten," observes Simonyi of Microsoft.

Fisher sees tragedy of a different kind. "The real price we pay is that as a specialist in any software technology you cannot capture your special capability in a product. If you can't do that, you basically can't be a specialist." Not that some haven't tried. Before moving to NIST last year, Fisher founded and served as CEO of Incremental Systems. "We were truly world-class in three of the component technologies that go into compilers but were not as good in the other seven or so," he states. "But we found that there was no practical way of selling compiler components; we had to sell entire compilers."

So now he is doing something about that. In April, NIST announced that it was creating an Advanced Technology Program to help engender a market for component-based software. As head of the program, Fisher will be distributing $150 million in research grants to software companies willing to attack the technical obstacles that currently make software parts impractical.

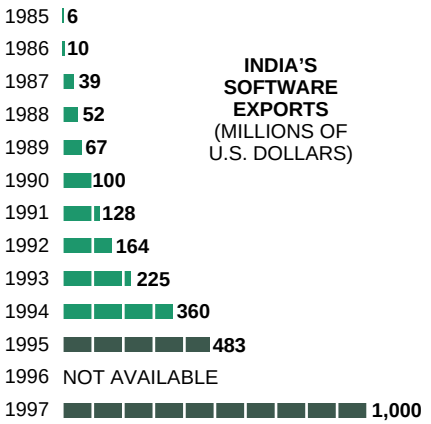
The biggest challenge is to find ways of cutting the ties that inherently bind programs to specific computers and to other programs. Researchers are investigating several promising approaches, including a common language that could be used to describe software parts, programs that reshape components to match any environment, and components that have lots of optional features a user can turn on or off.

Fisher favors the idea that components should be synthesized on the fly. Programmers would "basically capture how to do it rather than actually doing it," producing a recipe that any computer could understand. "Then when you want to assemble two components, you would take this recipe and derive compatible versions by adding additional elements to their interfaces. The whole

Bangalore building software for its Iridium satellite network.

Offshore development certainly costs less than body shopping, and not merely because of saved airfare. "Thanks to the time differences between India and the U.S., Indian software developers can act the elves and the shoemaker," working overnight on changes requested by managers the previous day, notes Richard Heeks, who studies Asian computer industries at the University of Manchester in England.

Price is not everything. Most Eastern nations are still weak in design and management skills. "The U.S. still has the best system architects in the world," boasts Bill Curtis of the Software Engineering Institute. "At large systems, nobody touches us." But when it comes to just writing program code, the American hegemony may be drawing to a close.



SOURCES: NIIT, NASSCOM

thing would be automated," he explains.

Even with a $150-million incentive and market pressures forcing companies to find cheaper ways of producing software, an industrial revolution in software is not imminent. "We expect to see only isolated examples of these technologies in five to seven years—and we may not succeed technically either," Fisher hedges. Even when the technology is ready, components will find few takers unless they can be made cost-effective. And the cost of software parts will depend less on the technology involved than on the kind of market that arises to produce and consume them.

Brad Cox, like Fisher, once ran a software component company and found it hard going. He believes he has figured out the problem—and its solution. Cox's firm tried to sell low-level program parts analogous to computer chips. "What's different between software ICs [integrated circuits] and silicon ICs is that silicon ICs are made of atoms, so they abide by conservation of mass, and people therefore know how to buy and sell them robustly," he says. "But this interchange process that is at the core of all commerce just does not work for things that can be copied in nanoseconds." When Cox tried selling the parts his programmers had created, he found that the price the market would bear was far too low for him to recover the costs of development.

The reasons were twofold. First, recasting the component by hand for each customer was time-consuming; NIST hopes to clear this barrier with its Advanced Technology Program. The other factor was not so much technical as cultural: buyers want to pay for a component once and make copies for free.

"The music industry has had about a century of experience with this very problem," Cox observes. "They used to sell tangible goods like piano rolls and sheet music, and then radio and television came along and knocked all that into a cocked hat." Music companies adapted to broadcasting by setting up agencies to collect royalties every time a song is aired and to funnel the money back to the artists and producers.

Cox suggests similarly charging users each time they use a software component. "In fact," he says, "that model could work for software even more easily than for music, thanks to the infrastructure advantages that computers and communications give us. Record players don't have high-speed network links in them to report usage, but our computers do."

Or will, at least. Looking ahead to the time when nearly all computers are connected, Cox envisions distributing software of all kinds via networks that link component producers, end users and financial institutions. "It's analogous to a credit-card operation but with tentacles that reach into PCs," he says. Although that may sound ominous to some, Cox argues that "the Internet now is more like a garbage dump than a farmer's market. We need a national infrastructure that can support the distribution of everything from Grandma's cookie recipe to Apple's window managers to Addison-Wesley's electronic books." Recognizing the enormity of the cultural shift he is proposing, Cox expects to press his cause for years to come through the Coalition for Electronic Markets, of which he is president.

The combination of industrial process control, advanced technological tools and interchangeable parts promises to transform not only how programming is done but also who does it. Many of the experts who convened at Hedsor Park agreed with Belady that "in the future, professional people in most fields will use programming as a tool, but they won't call themselves programmers or think of themselves as spending their time programming. They will think they are doing architecture, or traffic planning or film making."

That possibility begs the question of who is qualified to build important systems. Today anyone can bill herself as a software engineer. "But when you have 100 million user-programmers, frequently they will be doing things that are life critical—building applications that fill prescriptions, for example," notes Barry W. Boehm, director of the Center for Software Engineering at the University of Southern California. Boehm is one of an increasing number who suggest certifying software engineers, as is done in other engineering fields.

Of course, certification helps only if programmers are properly trained to begin with. Currently only 28 universities offer graduate programs in software engineering; five years ago there were just 10. None offer undergraduate degrees. Even academics such as Shaw, DeMillo and Basili agree that computer science curricula generally provide poor preparation for industrial software development. "Basic things like designing code inspections, producing user documentation and maintaining aging software are not covered in academia," Capers Jones laments.

Engineers, the infantry of every industrial revolution, do not spontaneously generate. They are trained out of the bad habits developed by the craftsmen that preceded them. Until the lessons of computer science inculcate a desire not merely to build better things but also to build things better, the best we can expect is that software development will undergo a slow, and probably painful, industrial evolution.

FURTHER READING

ENCYCLOPEDIA OF SOFTWARE ENGINEERING. Edited by John J. Marciniak. John Wiley & Sons, 1994.

SOFTWARE 2000: A VIEW OF THE FUTURE Edited by Brian Randell, Gill Ringland and Bill Wulf. ICL and the Commission of European Communities, 1994.

FORMAL METHODS: A VIRTUAL LIBRARY. Jonathan Bowen. Available in hypertext on the World Wide Web as http://www.comlab.ox.ac.uk/archive/formal-methods.html

Turning Green

Shell International projects a renewable energy future

JEFFREY HAMILTON *Liaison International*

WINDMILLS dominate the skyline on hills in northern California. Renewable energy sources such as wind power and photovoltaics may come to dominate world supplies after 2050, according to new analyses being performed by Shell.

Prediction is always dangerous, and predicting the fortunes of energy sources is the riskiest form of this professional sport. Still, after many years in the field Shell has a better track record than most. The giant corporation's planning group is credited, for example, with alerting the company's management to the possibility of an oil crisis before the oil price hike of 1973. So when Shell talks (particularly when it talks to itself), everyone tries to listen. At the moment, knowledgeable ears are trained in the direction of the Shell International Petroleum Company in London, a service company for the Shell group.

And what they are hearing is definitely not orthodox stuff. Shell's business environment group, headed by Roger Rainbow, has sketched a future in which renewable sources will grow to dominate world energy production by the year 2050. That perspective contrasts sharply with conservative studies by the World Energy Council (WEC), an international energy industry organization, and the International Energy Agency (IEA), an intergovernmental body. The WEC, for example, considers that "new" renewable sources, which include solar, wind, small hydroelectric, modern biomass and ocean sources, may account for only 5 percent of the world's energy output in 2020. In this view, fossil fuels will provide most of global energy needs through the middle of the next century, while nuclear fission plays an important supporting role.

The WEC's projections, the result of a three-year, $5-million study, were published last year. According to one of its midrange projections, annual global energy production will increase by 80 percent by 2020, to the equivalent of 16 billion metric tons of oil. That output will be needed to meet the needs of a human population that will be on its way from 5.5 billion (the 1990 figure) to 8.1 billion in 2020. (The numbers come from estimates by the United Nations and the World Bank.) The WEC and IEA studies presume that the new technologies will simply not have matured enough to capture a large fraction of the markets for coal, oil and natural gas.

But Rainbow and his colleagues, notably Georges DuPont-Roc, head of the planning group's energy division, disagree. Although the Shell exercise is not yet complete, Rainbow and DuPont-Roc have given officials at the World Bank and the U.S. Department of Energy a peek at the work in progress. And Peter Kassler of the Shell group has described some of the project's key aspects to the World Petroleum Council.

Kassler describes two possible geopolitical scenarios for the next 25 years. In one, the global trend toward economic liberalization and democratic reform in the 1980s continues to roll forward. That leads to a large increase in energy demand in developing countries, especially China and India, the world's most populous nations. At the same time, however, energy efficiency improves because of increased competition. Energy taxes internalize environmental costs, which help to stimulate the development of cleaner technologies.

Renewables gain importance in the second scenario, too, but less so, and in a distinctly grim setting. Regional economic and political tensions dominate the globe. Demand for oil increases, albeit slowly, but there is far less improvement in energy efficiency than in the first scheme. Protectionist policy and law weaken market forces. Oil price shocks exacerbate deteriorating international relations, and environmental anxieties spur government control of energy industries to ever stricter levels. New markets for renewables are "largely in poor countries" or are developed locally and cheaply.

Kassler points out that the protectionist option is bad news not only for oil companies but also for the environment. Under any plausible view, developing countries account for most of the growth in demand over the next 30 years. If they do not gain access to new, energy-efficient technologies, they will follow the energy-inefficient path taken by the developed countries.

Under the more optimistic view, Kassler speculates, renewable energy technologies may well "start to be competitive with fossil fuels around 2020 or

2030." He points out that fossil fuel technologies will probably be unable to lower costs as quickly as will the younger upstarts. Then "potential uses of the new technologies would grow." Developing countries might leapfrog over the industrialized nations toward an energy-efficient future.

Kassler foresees changes on the demand side, too. Virtual reality might, he conjectures, lead to a reduction in the demand for travel, thus breaking the long-standing exponential growth in personal mobility. In any event, fossil fuel use would start to decline around the middle of next century.

Emissions of carbon dioxide, which most atmospheric scientists expect to lead to significant global warming, would start to fall. In the midrange futures that the WEC considers, carbon dioxide concentrations will continue to rise from the current level of about 358 parts per million throughout next century, reaching about 600 parts per million in 2100, while still rising.

Rainbow says he is convinced that, for the long term, business-as-usual scenarios are "fundamentally and deeply flawed." He believes it is "obvious" that "there won't be that much coal, oil and gas being used in 100 years." Shell has considered "green" energy futures in the past, but they assumed stringent environmental regulation. The new work is remarkable because it envisions a sustainable future without draconian controls, says Christopher Flavin, an energy analyst at the Worldwatch Institute in Washington, D.C.

Not everyone is convinced that Shell has got the future right. Lee Schipper, an energy researcher at Lawrence Berkeley Laboratory, notes that the company deliberately considers wide-ranging possibilities. Schipper, who was himself formerly in Shell's planning group, indicates that the new analysis is "not yet good enough to go on the record." Rainbow says he expects to air more details about his group's thinking later this year.

Even in its embryonic state the DuPont-Roc/Rainbow vision is applauded by environmentalists such as Flavin. He suggests that Shell's conclusions "discredit" the conservative World Energy Council predictions. Flavin believes that hydrogen generated by electrolyzing water using power from photovoltaic plants will be the fuel of the second half of next century.

Nobody will predict the unfolding reality closely. But when a major oil company effectively projects the end of the fossil fuel age, it is a sure harbinger that we are moving into the future on fast-forward. —*Tim Beardsley*

Binary Disinfectants

Endowing computers with a software immune response

Every day two or three new computer viruses get written. From 10 to 15 percent of the more than 2,500 viruses ever authored have reached "the wild," a computer expert's term of art for that vast undefined electronic territory where a digital pathogen sets about infecting other computers. That number is enough to keep busy a small flourishing industry that writes commercial software to root out these digital microbes. Teams of experts at these firms analyze viruses with names like Junkie, Michelangelo or Jerusalem for hours or days to elicit signatures of a virus. One such company is even traded on the over-the-counter stock market.

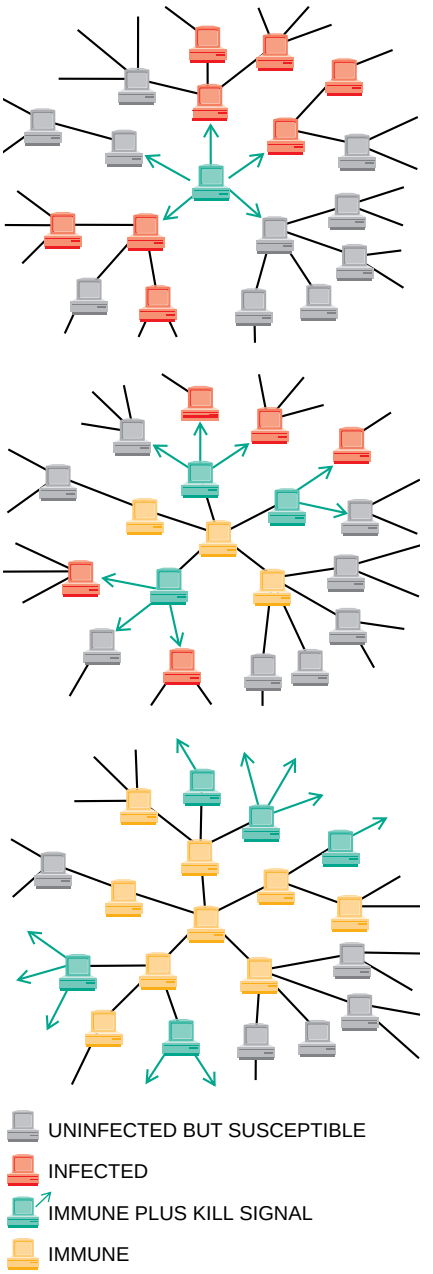
Viral detection programs, though, deserve further automation. They typically consist of several modules that check for suspicious system activity or changes to files. The software also contains a procedure to restore damaged files. Another vital function is a scanner that searches a computer memory or disk storage space for viruses by looking for a characteristic pattern of bytes, called a signature. But if the signature is not in the software's database the program is likely to miss the virus. So computer users must periodically get these programs updated. Having the antivirus software deduce a signature on its own has become a priority for a few commercial software companies.

In early July an investigator from the IBM Thomas J. Watson Research Center presented a technical paper on an approach to machine detection of viruses that may constitute an important advance. The researcher, Jeffrey O. Kephart of IBM's High Integrity Computing Laboratory, elucidated his ideas in a conference paper, "A Biologically Inspired Immune System for Computers."

Kephart compares the growing challenge of fighting computer viruses to the hopeless situation the Centers for Disease Control and Prevention would be in if it had to find a cure for each new strain of the common cold. He therefore suggests that protective measures should mimic the extraordinary capabilities of the human immune system to ferret out and destroy pathogenic organisms on its own.

A key element in the IBM software is the equivalent of a macrophage. In an animal, this kind of cell captures an invading organism, breaks up an antigen and presents a piece of it. Other immune cells use this fragment as a marker that helps them identify other appearances of the same microbe.

The software macrophage consists of a "decoy" program designed to make itself easy prey to a virus. One means of attracting infection is to interact frequently with the computer's operating system. By reading, writing and copying files, a program—in effect, a cyber immune cell—travels in and out of mem-



KILL SIGNAL, which warns of a virus and supplies repair information, is sent to neighbors by an infected computer that has already immunized itself against future infection. If uninfected, an adjacent machine is simply immunized; if contaminated, a computer is immunized and then sends a kill signal to neighbors.

ory the way a macrophage roams the body in search of foreign microorganisms. Other programs in the IBM immune system inspect the decoys to determine if they have been modified and thus may be infected.

Although a few other antiviral software companies also employ decoys, Kephart claims that a statistical analysis program to discern a viral signature—and a tool to inspect the way that the virus attaches itself to data or programs—is unique to the IBM software. The signature detector, he says, is able to identify most viral signatures better than a human could.

Right now IBM uses its automatic immune system as capital equipment, a tool to troubleshoot customer problems and update protective software sold to customers. "This [software immune system] enables us to keep pace with the influx of new viruses with just one human virus expert who analyzes viruses half-time, as opposed to the dozen or more virus analysts employed by some other antivirus software vendors," Kephart writes. Only the most elaborately crafted binary germs still require expert eyeballing.

So far computer virus epidemics have been rare. Most contagions remain localized; transmission from machine to machine usually occurs by passing floppy disks. But the growth of office networks and the Internet is expanding the vectors through which viruses can infect a vast number of computers.

In response, IBM researchers have devised a mass vaccination and treatment strategy. Notification of infection is sent to neighboring machines on a network. Besides alerting other machines, this "kill signal" also contains information about a virus's signature and provides repair instructions. The recipient, in turn, notifies other machines if it finds itself infected. "The signal follows the path of a virus and stops it before it gets very far," Kephart says. A similar approach, he notes, was taken during the campaign to eradicate smallpox. Vaccinations went to those who had come in contact with someone infected.

Since 1991 various elements of this computer immune system have been used to extract the signatures of more than 2,500 different viruses that work on IBM PCs. IBM now plans to market its entire antiviral repair suite. The software would probably reside on a single computer in an office network that would keep a vigil for disease-causing code. Antivirus program updates would become rarer, although there would be a need for occasional revisions for the most pesky viruses that escape the scrutiny of the automatic immune system.

For Kephart the connection between silicon and biological immunology is more than metaphor deep. At the end of his paper, he remarks perhaps only half in jest that the field of digital virology may offer employment opportunities for theoretical immunologists. Their insights may be needed, because even the most isolated corners of the world are not immune from infection. In January a virus called *barrote,* which informally means "jail" in Spanish, was discovered at Spanish and Argentine scientific outposts in Antarctica. Immunodeficiency is more than a human condition. —*Gary Stix*

Heat and Light

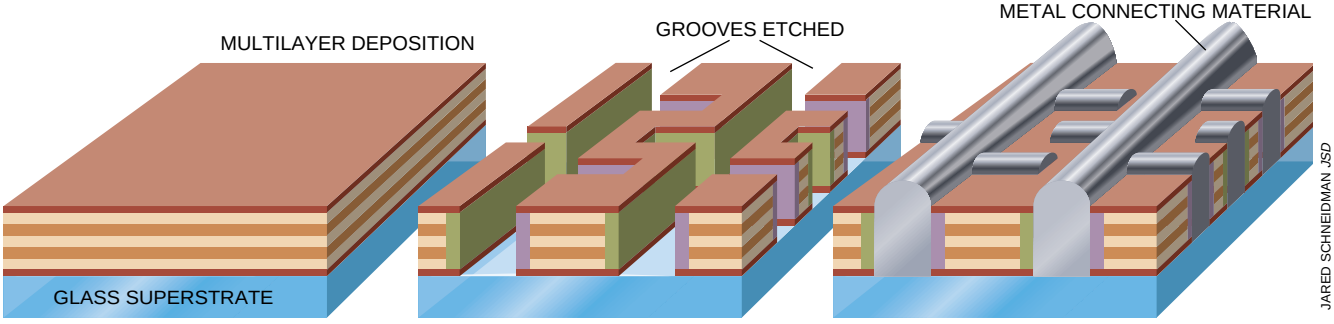
As titans battle, a midget attempts to steal the march

At the turn of the year William H. White, the U.S. deputy secretary of energy, made an announcement of the kind about which federal technocrats dream: the successful completion of an inexpensive (by Washington standards) development project, whose consequences for competitiveness and the environment could be profound. It was a $6.26-million effort to produce an efficient, low-cost photovoltaic cell. United Solar Systems Corporation in Troy, Mich., conducted the research and funded half its cost. The Department of Energy footed the rest of the bill. That was then. Today lawsuits and countersuits ensnare the project, accusations blister corporate reputations and the entire struggle may be rendered moot by a band of university technologists from Down Under.

Solarex, an Amoco-owned firm in Frederick, Md., fired the first legal gun. Solarex alleges that its fundamental patents on amorphous silicon technology are infringed by the United Solar device, which uses amorphous silicon and two different alloys of germanium and silicon. The cell achieves more than 10 percent efficiency at a cost substantially lower than that of existing designs. United Solar is a joint venture of Energy Conversion Devices and Canon of Japan. David Carlson, who is now a vice president of Solarex and did the original research on amorphous silicon in the early 1970s, says Amoco decided to sue.

United Solar has now countersued, alleging infringement of its patents by Solarex, so the stage is set for a protracted legal struggle. Subhendu Guha, the inventor of the United Solar cell, insists that his design is the result of more than 15 years of independent research. "Solarex, the controlled subsidiary of a major oil company, Amoco," Guha states, "is engaged in a campaign of bringing patent infringement suits trying to close down amorphous silicon solar cell manufacturers who are committed to solving the energy crisis of this and the next century."

In fact Solarex has successfully sued another company that was using amorphous silicon technology, Siemens Solar Industries (at the time Arco Solar), and it has a case pending against Advanced Photovoltaic Systems of Princeton, N.J. If Solarex is successful against United Solar, it would prevent the company from manufacturing the low-cost cell at a factory it is building in Newport News, Va.



PHOTOVOLTAIC CELL devised by Australian researchers uses alternating layers of "p-" and "n-type" silicon on glass. A laser etches grooves, which are then filled with metal for contacts. Some grooves connect to p-type silicon, others to n-type.

The matter also raises the question of whether the DOE has adequately protected the public interest in the research projects that it supports, observes James Caldwell, technical director of the Center for Energy Efficiency and Renewable Technology in Sacramento, Calif. Caldwell, who was president of Arco Solar when it was sued by Solarex, notes that, like United Solar, Solarex has received money from the DOE (it will not say how much) to develop amorphous silicon technology. Yet for applications that demand high power, most of the corporation's products use a different form of silicon.

Several Japanese manufacturers nevertheless use amorphous silicon in consumer products through an agreement with RCA Corp., which is where Carlson first developed the substance. Caldwell wonders whether the DOE should not try harder to ensure that U.S. companies are able to exploit valuable inventions arising from research on energy sources that it supports. Solarex, which has a large share of the U.S. market for photovoltaics, felt threatened by United Solar Systems's design, according to a source at Solarex.

The outcome of the battle might be moot if a cell invented by Martin Green, Alistair B. Sproul and their associates at the University of New South Wales in Sydney, Australia, is as successful as the researchers expect. Green's group holds the world record for efficiency of solar cells in laboratory rigs. Moreover, one of Green's innovations, buried contacts, is already in use in panels made by B. P. Solar in Australia. Green has licensed the technology to Solarex. Buried contacts, which are metal connectors recessed into the cell in grooves cut by a laser, eliminate the need for wires lying on top of the cell that block out some sunlight.

Green's cell consists of many very thin layers of silicon, doped so that "*p*-type" layers alternate with "*n*-type" layers. Buried contacts made the development possible: traditional methods of attaching contacts could not be used to connect a wire to all the even-numbered layers (for example) in the multilayer sandwich that is the cell design. Sproul says he believes the new cell should be able to achieve 25 percent efficiency at a cost of only $1 per square foot. That would make solar power competitive with electricity from fossil fuels in many homes. But Sproul estimates that 10 years of development work will be needed to turn his laboratory-scale rig into a commercial product. At least the germination process Down Under should be free of the legal tangles impeding progress to the north. *—Tim Beardsley*

Food Fights

Is it a drug or a carrot stick?

The new, it has been said, arrives disguised as the old. And sometimes it is just the reverse. The notion of food as elixir, a hand-me-down from antiquity, has reemerged bearing a new set of names; among them are nutraceuticals, designer foods and functional foods. By whatever name, the health benefits of sulforaphane from broccoli, beta carotene from carrots or calcium from a variety of sources have caught the attention of corporate research and marketing departments at major drug companies and consumer food products concerns.

"This is potentially a huge new market," says Rick Guardia, group vice president of technology for Kraft General Foods. The once immortal baby-boom generation has worries about the prospect of the chronic diseases that come with aging. Facing price pressure, drug companies want to embrace preventive instead of curative medicine. Businesses now give employees incentives to watch blood pressure and cholesterol.

Such opportunities could begin to blur the distinction between a food and a drug company. At Merck's Kelco division, an executive now bears the word "nutraceutical" in a job title. Procter & Gamble is setting up a group in its health care sector to probe the possibility of a new line of business. Pfizer is securing the rights for natural sources of beta carotene.

The industrial activity is complemented in academia. The University of Illinois has a functional foods program; Tufts University is trying to organize one. A steady flow of findings is being reported in a number of medical journals (some of the studies funded from corporate coffers).

Pressure has begun to build on the regulatory authorities. Industry wants latitude to perform clinical tests and then summarize the study results on the labels of chicken soup cans or fruit-drink containers. It believes that the Food and Drug Administration has shown an overcautious attitude toward letting companies convey information about foods or their components that have been ingested for hundreds of years without ill effect. "I have argued very strenuously that this FDA's restrictive interpretation amounts to nothing short of censorship," says Peter Barton Hutt, a Washington attorney and professor of food and drug law at Harvard Law School. "The only people who are

Power Medicine

At least five medical device manufacturers have research or clinical trials in progress to bring to market a transdermal patch supplemented with a battery and electrodes that supply a slight current to the skin. In the U.S. the unelectrified patch has been used for delivering seven or eight compounds—nicotine for repentant smokers is one; scopolamine for travel sickness is another. The drugs transported through the skin have a molecular weight of no more than a few hundred daltons. With an electrical patch, a weightier molecule—an ionized peptide drug that inhibits bone loss, for example—moves through the top layer of skin with the flow of current.

The typical electrical patch is equipped with a battery, two electrodes and a microprocessor-based controller. One developer, an Irish company called Elan, even has an agreement with the Swiss manufacturer of Swatches to design a wristwatchlike drug-delivery device.

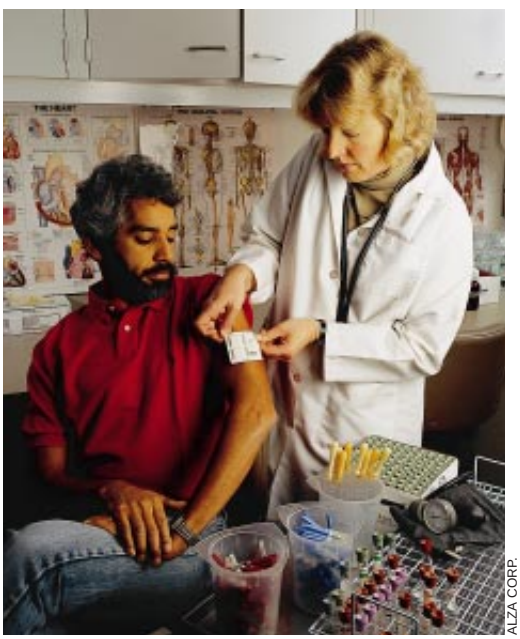
Manufacturers believe the electrical patch could broaden the market for transdermal devices. The patch may also be the ideal way to infuse the exotic peptides and proteins devised by biotechnology companies.

Safety is of course an essential feature. The major developers—Becton, Dickinson & Co., Iomed, Alza and Elan—see the patch as a relatively foolproof means by which patients can administer painkillers to themselves. The controlled level of electric current ensures a more predictable level of drug delivery than does the patch alone. The precise amounts delivered could avoid the overdosing with opioids that has occurred with the nonelectrical patch. A patch equipped with a button could enable patients to administer a painkilling drug as needed. This technology is cheaper and gives a patient more mobility than do current self-medication devices that inject a drug with a pump and needle.

Other research has shown how the electrical patch could be used to administer anticancer agents, including drugs for treating some forms of skin cancer. The patch might also infuse luteinizing hormone to increase fertility, or calcitonin for osteoporosis and antibiotics to fight infection of burned tissue.

In the competition to reach the market, manufacturers have devised new variants on the original design. Instead of applying a steady, low direct current, the patch developed by Cygnus Therapeutic Systems delivers a millisecond pulse of a few hundred volts to generate a high electrical field. The technique, called electroporation, is said to increase skin permeability, thereby enabling a drug to infuse more readily.

Any new product will have to be evaluated carefully to avoid irritation caused by stimulating the skin with a current. Food and Drug Administration approval could prove a challenge: regulators may struggle in doing a technical assessment of an item that is half drug, half medical device.



ALZA CORP.

NO SHOT IN THE ARM is needed with an electrical patch that can infuse into the skin larger molecules than can conventional transdermal patches.

If they succeed, manufacturers have already conceived of a late-model electrical patch. Electric current from a battery—and the consequent flow of drug—could be adjusted by incorporating minute sensors into the patch's electronic control unit. With the smart electrical patch, a physician may one day be able to say, "You won't feel a thing"—and really mean it. *—Gary Stix*

not permitted to talk about food are the people who make the product."

The FDA is unimpressed. The agency is responsible for the safety of substances ingested by the public every day, not for nurturing new markets. The Nutrition Labeling and Education Act of 1990 (NLEA), which began to take effect only last year, does allow the use of general health claims on food labels: the link between the sodium in salt and hypertension, for example.

Broad assertions about low sodium or the healthy aspects of consuming fruits and vegetables can be made by any company that meets the FDA rules. But food and drug manufacturers want to differentiate one product from another by making claims based on research carried out on a specific substance—a patented form of calcium that is more readily absorbed in the body, for example. The industry argues that the NLEA sets no clear regulatory pathway for justifying an exclusive health contention. "Without those rules, I can't go to management and ask for $10 million to prove how garlic might prevent the common cold," Guardia says.

The inventor of the word "nutraceutical," Stephen L. DeFelice, a former chief of clinical pharmacology at Walter Reed Army Institute of Research, has proposed the establishment of a panel of experts that would be independent of the FDA. This nutraceutical commission would evaluate the merit of research findings conducted by a company. If valid, the company would be given exclusive rights to make a health claim for seven years.

The process for approving a nutraceutical would resemble that used to grant exclusive marketing status to pharmaceuticals for rare diseases, which are called orphan drugs. DeFelice's ideas have not been universally endorsed by industry—companies do not want another regulatory body.

The FDA is not about to readily embrace an "orphan" carrot-juice cocktail. Agency officials say that justification for their conservative wait-and-see stance is bolstered by recent experience. The NLEA requires significant agreement among scientific experts on whether a health benefit can be attributed to a particular class of food. During the past year the agency was considering whether to reevaluate its earlier decision not to allow a general health claim for antioxidant compounds, such as beta carotene, which some studies suggest lower the risk of heart disease and cancer.

Then this spring a study by the National Cancer Institute and Finland's National Public Health Institute showed that Finnish men who were heavy smok-

ers and ingested beta carotene had higher rates of cancer than did those smokers who just took a placebo. "The most carefully controlled study that has ever been done came in, and it blew the antioxidants completely out of the water," says Fred R. Shank, the FDA's director of the center for food safety and applied nutrition.

The agency's defenders emphasize that a process already exists for tying a claim to an individual product. If a nutraceutical is a cross between a food and a drug, apply the strictest standard: consider the compound to be a drug. "What some companies would like to do is make pharmaceutical claims about food and disease without going through the process of proving their pharmaceutical effectiveness," says Mark Silbergeld, director of Consumers Union's Washington office.

Foreign markets offer more freedom. Procter & Gamble has licensed to a Japanese company the rights to manufacture calcium citrate malate, a form of the mineral that it contends is quite easily absorbed by the body and has been shown to promote bone growth in children. Takara Shuzo Company recently began marketing a fruit-flavored drink enhanced with this formulation; the label contends that the product helps to increase bone mass.

The Japanese Ministry of Health and Welfare gave Takara permission to use that label after the company submitted data from studies on bone density that Procter & Gamble had helped fund at universities in the U.S. and that were then published in the *New England Journal of Medicine*. For several years, the calcium health claim will be allowed only for this compound.

In the U.S., the FDA allows a label on any product that meets its rules to say that calcium helps to reduce the risk of osteoporosis. So Procter & Gamble's calcium-enriched Hawaiian Punch product in U.S. markets is just one more calcium product on the shelf.

Over time—probably a longer stretch of it than the industry would wish—a new regulatory classification for the health-giving properties of an individual food will probably emerge in the U.S. The FDA's Shank believes a revamped process will eventually be worked out for making health claims about food, perhaps even one akin to the current means for approving orphan drugs. In its slow and deliberate manner, the FDA may move toward becoming the Foods as Drugs Administration. "The focus," Shank says, "is shifting from what provides normal growth and development in people to what provides optimal health."

—Gary Stix

Crabshoot

Manufacturers gamble on cancer vaccines—again

The idea of using a vaccine for stimulating the immune system to fight tumors has a long, undistinguished history. Because early attempts were based on extracts of tumors, which are inherently variable, occasional reports of success generally led nowhere.

In the past few years the climate has radically changed. Advanced cell culture techniques have made possible the identification of specific antigens—substances recognized by the immune system—that are present on tumor cells in larger amounts than they are on ordinary cells. Using monoclonal antibodies, workers have been able to distinguish different types of cells in the immune system from one another so that the body's responses can be minutely monitored. More recently, tools such as the polymerase chain reaction (PCR) have been used to analyze and copy tiny samples of gene fragments. Rapid progress in identifying cytokines, the chemicals that control the immune system, raises the hope that essential support from *T* cells can be mobilized.

As a result, work on therapeutic vaccines—for infectious diseases as well as cancer—is enjoying a high-tech renaissance. At least 16 biotechnology and pharmaceutical companies around the world, including some of the biggest names, have research programs in vaccine therapy for cancer. Several have clinical trials already under way, and many more plan to initiate trials in the near future.

Cytel in San Diego is betting that small is beautiful. Cytel's approach is to analyze published sequences of tumor-associated antigens in order to find in them characteristic peptides that might bind to histocompatibility complexes. The rationale is that histocompatibility complexes on cells would present such peptides to *T* cells. The company tests selected candidates for their ability to stimulate human *T* cell activity. Cytel has a pilot study in progress with Steven Rosenberg, chief of surgery at the National Cancer Institute (NCI). In the study, killer *T* cells from patients with a variety of tumor types are cultured outside the body and exposed to stimulatory peptides before being reinfused.

Lynn E. Spitler, president of Jenner Technologies in Tiburon, Calif., says she fears peptides might be too small to elicit immune responses adequate to fight cancer. Jenner is actively developing two vaccines that employ much larger proteins. The proteins incorporate multiple sites that Jenner has shown can excite *T* cells. An adjuvant—a cocktail of immune-stimulating chemicals—is a key part of the formula. Jenner's targets are prostate, colorectal and lung cancers.

MedImmune in Gaithersburg, Md., is, like Cytel, interested in provoking immune reaction against tumors by means of short peptides. The company generates the peptides by splicing genes into BCG, a harmless bacterium widely used as a preventive vaccine for tuberculosis. The resultant engineered bacteria express the peptides on their membranes, thus stimulating *T* cell activity, says Scott Koenig of MedImmune. The company is investigating antigens involved in breast and colon cancer as well as in melanomas.

Some companies have decided that antigens are most effectively presented to the immune system by live viruses. Cantab Pharmaceuticals in Cambridge, England, has a pilot study in progress of a candidate therapy for cervical cancer. Cantab uses vaccinia virus—better known for preventing smallpox—to present antigens to patients' immune systems. The antigens are similar to ones that can stimulate immune responses against related cancers in animals. Genes encoding the antigens were identified by PCR and other techniques.

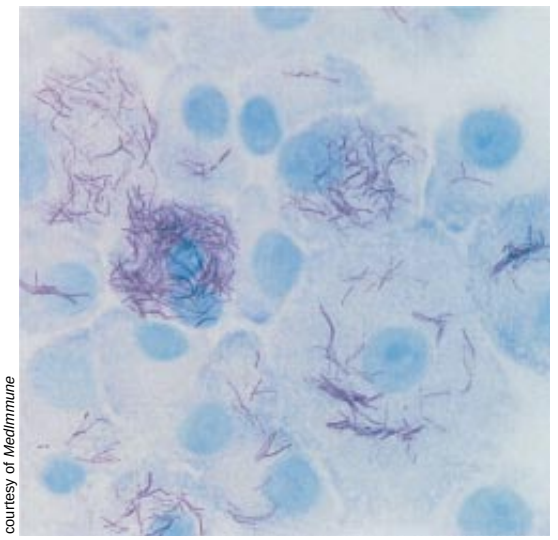
In the Massachusetts Cambridge, Therion Biologics has started clinical studies that test another use of vaccinia. It has engineered its virus to produce carcinoembryonic antigen, which is borne by many kinds of tumors. Therion also hopes to instill in its vaccinia the ability to make cytokines.

Not all antigens are proteins—the immune system can recognize sugars and other chemicals on cells as well. Biomira in Edmonton, Alberta, has developed a vaccine for breast cancer that mimics an unusual form of sugar found on the malignant cells. B. Michael Longenecker of Biomira says that in a pilot study designed to test toxicity, four of 13 patients with metastatic breast cancer exhibited substantial reduction in tumor size, and six remained stable. ImClone Systems in New York City is testing monoclonal antibodies that are "anti-idiotypes"—that is, antibodies that resemble antigens. The advantage of this approach is that antibodies can be made to resemble antigens that are not proteins. ImClone's antibodies resemble gangliosides, molecules with sugar and lipid components that are found in large amounts on a variety of tumors.

Although many vaccine developers are boutique operators, major pharmaceutical corporations are also coming into the field. Boehringer Ingelheim Pharmaceuticals and SmithKline Beecham have initiated research. And Chiron Corporation in Emeryville, Calif., one of the oldest of the new generation of biotech companies, has initiated a collaboration with Thierry Boon of the Ludwig Institute for Cancer Research in Brussels to investigate melanoma-associated antigens in a vaccine. The antigens will be given with interleukin-2, which Chiron manufactures.

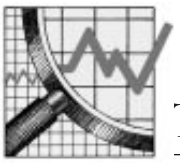
Chiron has also initiated a collaboration with Viagene in San Diego that combines vaccine therapy for melanoma with gene therapy. Viagene has a product that can insert the gene for gamma-interferon into tumor cells. Experiments suggest that tumor cells thus modified can present antigens to the immune system in such a way as to invite destruction. Somatix in Alameda, Calif., is pursuing similar research.

Mainstream cancer researchers are becoming more interested in cancer vaccines, although they are not yet sold on them. "We know a lot more about tumor antigens than we did," comments Bruce A. Chabner, head of cancer treatment at the NCI, "but I do not think I can say we're optimistic." Boon thinks there is a 50 percent chance that vaccines will become valued cancer therapies. "In two or three years," he adds, "we will know whether that has changed to 10 or 90 percent." —*Tim Beardsley*



courtesy of *MedImmune*

GENETICALLY ENGINEERED BCG bacteria* (red) *might stimulate cells of the human immune system* (blue) *to fight tumors.



Not Worth the Paper It's Printed On

A rich man died during the early days of Germany's Weimar Republic (so the story goes) and left all he had to his two sons. To the prudent one he willed his fortune, and to the profligate his collection of bottle caps. When the inflation came, the prudent son invested wisely and lost everything; the profligate sold the bottle caps and survived. Or, as Harvard University economist Jeffrey Sachs says, hyperinflation (price increases of more than 50 percent in one month) has "very costly, arbitrary distributive consequences."

When prices rise so quickly—sometimes even doubling in four days—they turn a nation's economy upside down and render banknotes so much wastepaper. The same disaster that overtook Germany, Austria, Hungary, Poland and Russia in the 1920s and China, Greece and Hungary in the 1940s attacked Argentina, Bolivia, Nicaragua, Peru and others in the 1980s. Brazil flirted with it earlier this year before imposing wage controls and a new economic stabilization plan (the sixth in 10 years) this summer. How does money lose its value so suddenly, and why are the effects so devastating? Who might fall next?

It is all very simple, according to Sachs, who helped Bolivia and Poland get out of the inflation trap. First, a government allows its expenditures to exceed its revenues dramatically. Then, instead of cutting outlays or collecting more revenue, it prints money to pay its bills. During the 1920s and 1940s, national bankruptcy came on the heels of world wars; in the 1980s, it was the aftermath of excess international lending.

Simply creating more money does not increase the amount of goods available in the country, and so inflation soaks up the excess currency. If the treasury doubles the supply of reichmarks or pengös or rubles, prices double. Each reichmark, pengö or ruble is worth half of what it was, but the government's need for gasoline, paper, bureaucrats, soldiers and so forth remains the same. So it prints more money to cover the now inflated prices. Worse yet, citizens quickly realize that money is not a good investment. They seek to get rid of it, and so, as with any unwanted commodity, its value drops further. The spiral soon becomes vertiginous.

By itself, the falling value of money is not ruinous. A number of countries, most notably Brazil and Argentina, learned to cope with inflation rates of several hundred percent a year by indexing wages, pensions, savings accounts, loans and other financial objects. Just as automobile workers or Social Security recipients in the U.S. received cost-of-living adjustments, the governments of high-inflation countries compensated their citizens. "Indexation can be very effective," says John Williamson of the Institute for International Economics—high and low rates of inflation are not qualitatively different.

Indeed, in theory a perfectly indexed economy could be almost indistinguishable from one with no inflation at all. Yet the practice is unpleasantly different. If holding banknotes or uncashed checks for even a day means losing money, then financial transactions of all kinds become more difficult. "Finance is a grease for the economy," states Alan Gelb of the World Bank. "It's an intermediate input for production, the same way that a railway system is." And meddling destructively with money will hurt an economy "just as much as if you screw the rail system up," he asserts.

People lose confidence in money, and they race to get rid of it before its worth drops any further.

Furthermore, as Sachs notes, most countries do not come on an episode of rampant inflation with indexing mechanisms in place. And although the principles of hyperinflation may be rigidly mathematical, its implementation depends on the psyches of millions of merchants, employers, customers and laborers. Each individual who loses confidence in the value of money—or becomes aware of just how far it has already been debased—starts the race to get rid of as much as possible before its worth drops any further.

As a result, Gelb points out, every increment of inflation brings with it an increase in uncertainty. Prices cannot rise continuously (imagine thousands of store clerks constantly scribbling or ringing up purchases with one eye on the clock), and so they leap upward in fits. Some merchants may readjust in the morning, others at night and others at noon; prices thus depend unpredictably on where an item is bought and when, rendering the notion of a unified market meaningless. And if there is no market, Gelb says, inflation indexes cease to represent anything real: "Prices might be rising at 50 percent a month, but when you go shopping, the increase could be half that or twice that."

Faced with such massive uncertainty, he comments, people avoid risk by abandoning the local currency for more stable alternatives such as dollars or deutsche marks. Capital flight adds fuel to the inflationary firestorm and further reduces economic activity.

Only rarely does a nation manage to wind down hyperinflation gradually, Williamson says. Instead the government collapses, and a new one capable of keeping its accounts in order takes its place. Often the replacement has been authoritarian, Sachs notes: a danger that looms in the economic crises in Russia, eastern Europe and elsewhere.

Surprisingly, hyperinflations stop even more abruptly than they start. Once the treasury's printing presses stop, according to Sachs, so does the crisis. By the time stabilization comes, he explains, almost everyone is doing business in dollars or other "hard" currencies, and so stabilizing the value of the peso, real or ruble is automatic once fiscal responsibility is restored. Indeed, inflation that stops short of complete monetary collapse may be more difficult to bring under control, Sachs observes—indexing institutionalizes rising prices. Finance ministries must carry off the most delicate footwork to give everyone the "last" increase.

In line with the symbolic nature of money, many nations (including Brazil for the fourth time this summer) issue a new currency to mark the end of their insolvency. Most economists agree, however, that the effect of such replacements is largely psychological. Indeed, replacing pounds with shekels or cruzeiros with reals while leaving underlying problems untouched is a sure way to lose the last shreds of economic credibility, Williamson remarks: "Sooner or later people are going to find out, and then you're lost." —*Paul Wallich*

A Subway Named Turing

The Tweedle twins and I were strap-hanging on the New York City subway. Delia Tweedle was universally known as Dee, so her brother had inevitably become Dum. Even though his real name was Seymour. As usual they were interrupting each other.

"Well, if the universe is algorithmic, then strong AI—"

"Don't be pedantic, Dee, what you mean is computers that think—"

"Must be possible in principle."

"Why?" I said.

"If our universe is algorithmic—"

"You could set up a computer to simulate it—"

"Which would therefore simulate everything in it, including us having this conversation," Dee concluded.

"You realize that if you're right, then a sufficiently complex subway system could become intelligent?" I said. "It would think rather *s-l-o-w-l-y*… but it would still be able to think."

"That's dumb," Dee exclaimed. "A subway can't think."

"Maybe not. But a subway can compute, according to a fascinating article I've just read in the latest issue of *Eureka*. It was written by Adam Chalcraft and Michael Greene, and it's about the computational abilities of train sets."

"You mean *toy* train sets? Rails and points and tunnels with sheep painted on their sides?"

"That's right, Dee. And whatever a train set can do, a subway surely can. It's not quite your hyperparallel superduperultracomputer, but in theoretical computing ability, equally as powerful. A computer is, after all, just a huge switching circuit with adaptable switches. And trains can switch tracks using points. What Chalcraft and Greene asked was: If you've got a big enough stock of straight and curved track, bridges and various kinds of points, but you've got only one engine and *no* rolling stock, then what computations can you do if you set up the right track layout?"

"I don't see how a train can compute at all," Dee puzzled. "It's just a thing on wheels that moves along the rails."

"Electrons are just things that move along wires, but they are what computers compute with," I pointed out. "In both cases, the computational aspect is a matter of interpretation. For a train layout, the idea is to encode an input as 0's and 1's corresponding to the settings of various points. Then you run the train through the layout, and sometimes those settings change, which in turn alters the path of the train. Eventually the train is sidetracked onto a line leading to a terminal, the program 'stops,' and you read the output from the settings of the same collection of points."

"Okay, I see that it might work. But does it?"

"Let me start with the simplest switching unit, known as a lazy point [*see top illustration on page 106*]. It's a Y-shaped piece of track. A train entering the Y from below runs up the upright and out of whichever arm of the Y the points are set for, leaving them unchanged. But a train that enters from one of the arms will—if necessary—reset the points so that they connect that arm to the upright and then exit via the upright. Lazy points have two states, depending on which arm is connected to the upright: call them left and right.

"The next type of point is a sprung one. It's like a lazy point except that any train entering along the upright of the Y always leaves by the same arm. The third kind of point is a flip-flop.

"With a flip-flop, trains always enter along the upright of the Y and exit through the left and right arms alternately," I went on. We lurched judderingly into the Liberty Avenue station. "The big question is: Given these components, can we build a computer?"

"What kind?" wondered Dee.

"A Turing machine," I said. "Alan M. Turing proved that his simple model of a computational system can do anything that a programmable digital computer can. Think of a Turing machine as a box that can travel along a very long tape of square cells, each containing either the symbol 0 or 1. You can either use an infinite tape or, if you don't like infinities, you just have to be prepared to add some more cells to the tape if you need them.

"The box can be in any of a finite set of internal states, depending on what hardware is inside it. For each combination of its own state and the digit written on the cell immediately below it, it must obey a small list of instructions, like this:

'Leave the current tape digit alone/change it.

'Then move one space left/right.

'Then go into some specified internal state ready for the next step.'

"Or the instruction can be just 'STOP,' and the computation then finishes."

"Give me an example."

"Okay. Here's a typical list of rules for a Turing machine with three states—1, 2 and 3:

'State 1, Digit 0: Change digit, move left, go to state 2.

'State 1, Digit 1: STOP.

'State 2, Digit 0: Leave digit, move right, go to state 3.

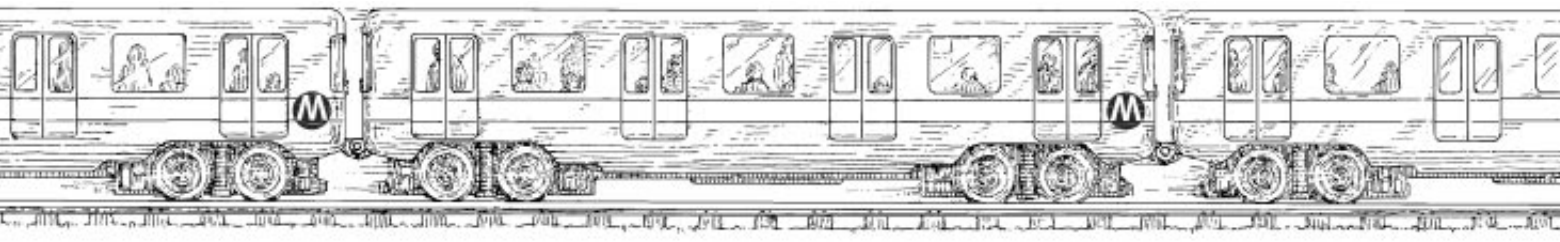
'State 2, Digit 1: Change digit, move right, go to state 2.

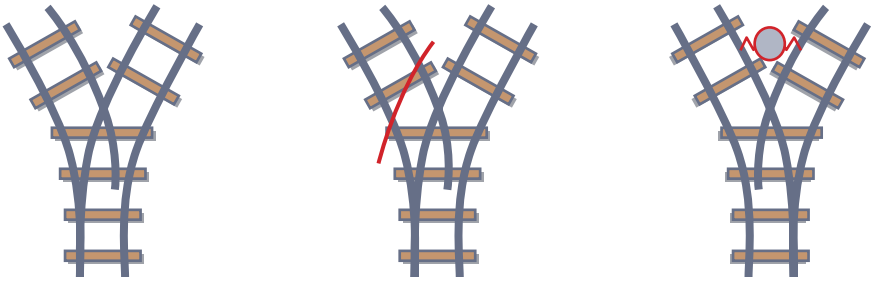
'State 3, Digit 0: Change digit, move right, go to state 1.

'State 3, Digit 1: Leave digit, move left, go to state 2.'"

"What does that compute?"

"I haven't the foggiest idea, Dum. Try it and see. Sensible programs generally need a lot more states than three, anyway. The digits on the tape provide the computer's input. The list of which instructions to perform for which internal state of the box forms the program, and the list of digits on the tape when





THREE TYPES OF POINTS

the computation stops is the output. Amazingly, these simple devices can carry out any algorithm whatsoever. So all we need is to find a train layout that simulates any chosen Turing machine."

"Tricky."

"Yes. It helps to break down the problem into a series of stages. Now, the idea is to find a train layout that can play the role of the box. Instead of using a tape, you just plug an enormous number of these boxes into each other, side by side, to represent the whole tape [*see illustration below*]. Each box will have several tracks coming in from the left and the same number going out the right—one track for each internal state."

"So instead of the box moving along the tape—" Dee began.

"The train moves along the row of boxes," Dum finished, enthusiastically.

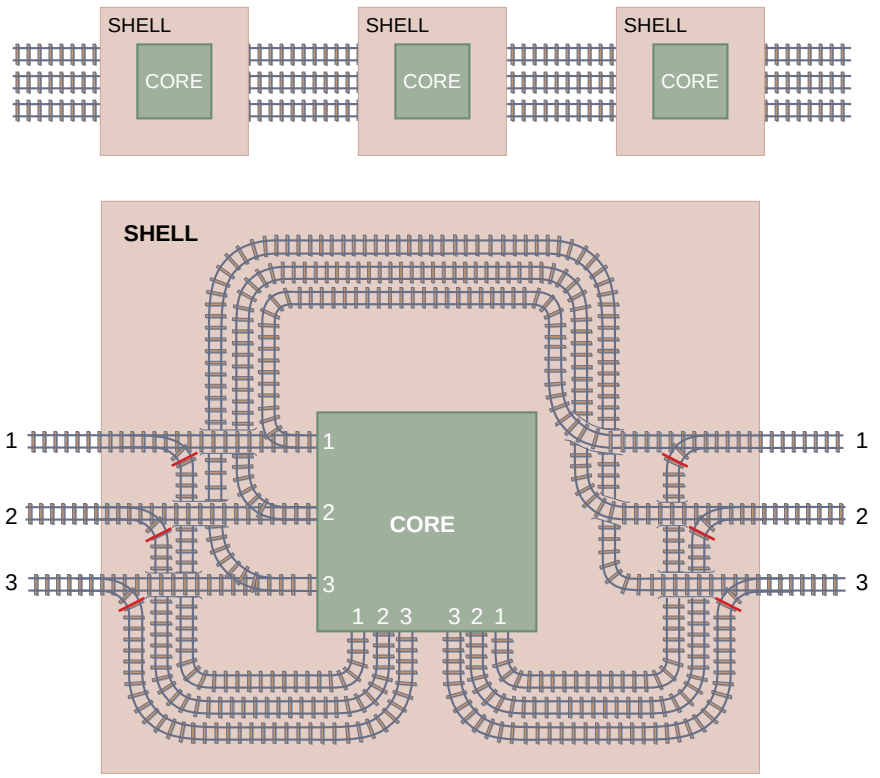
"You can tell which 'cell' of the 'tape' is being worked on—"

"By which box the train is in. Neat."

"Yeah," Dee agreed. "But what do you put in the box? Schrödinger's cat?"

"I'll explain how the box is designed in stages. The train tracks are used as both input and output lines, so the box doesn't 'remember' which direction the trains came in from. Therefore, it can be set up as an outer shell that feeds trains from both input lines the same way into a core and conducts them out again according to the Turing program. Then we can ignore the outer shell and concentrate on the design of the core."

"You're going to need—" Dum started.



TURING MACHINE TAPE can be replaced by series of identical boxes **(top).** ***Each box*** **(bottom)** ***consists of a shell, which ensures that trains entering from either direction are treated alike, and of a core, which carries out the rules of the Turing machine.***

"Subroutines," Dee said. They'd been thinking ahead, as usual. A subroutine is a part of a program that can be used repeatedly by "calling" it from any other part. You can build complex programs by stringing together subroutines.

"Yes," I concurred. "You can set up a subroutine by hooking up a self-contained sublayout to a whole series of lazy points. Then the train comes in, setting the points as it does so, and wanders around the sublayout until it has carried out whatever subroutine that the sublayout computes. Finally, it exits by the same track that it came in on because of the way it set the points on entry. Using one lazy point for each input line, the trains can all be made to enter from the left, carry out the subroutine and exit to the right along the same track they entered on [*see top left illustration on opposite page*]."

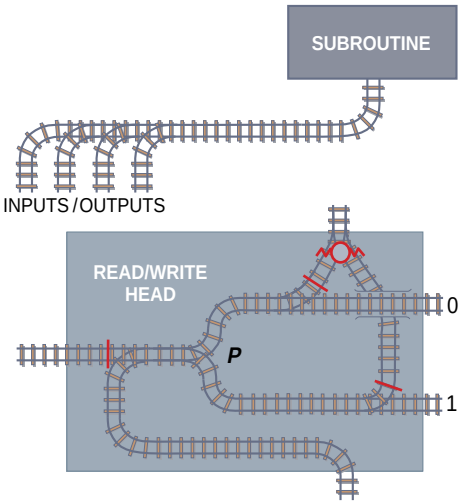
"Oh, right."

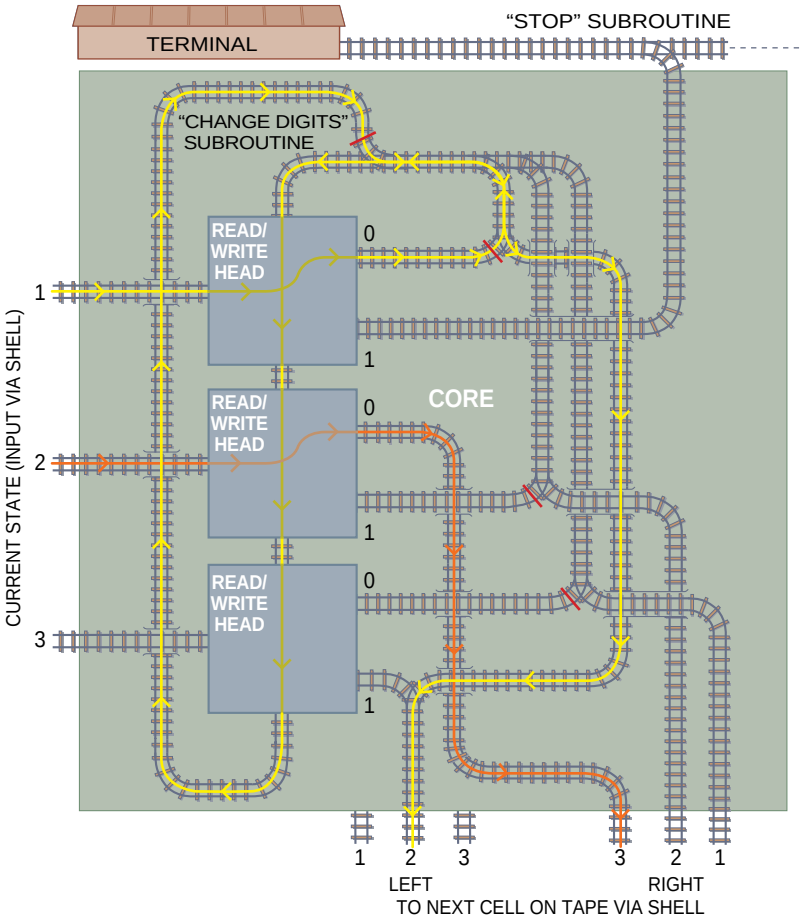
"Now, you need one more piece of gadgetry: a read/write head. If a train comes into a read/write head from the left, then it exits along line 0 or line 1 depending on the digit at the 'current' point of the tape. If a train comes in from above, then it swaps the 0 and the 1 and exits at the bottom. To achieve this, the lazy point *P* is set to redirect the train along output line 0 or 1 according to the digit on the 'tape' at that square. The flip-flop is set so that the first train entering from the top switches *P* to the other position.

"Having got all these bits and pieces, you build the inner core of the box [*see right illustration on opposite page*]. You may need some bridges to avoid the tracks crossing, but we can ignore that. The core consists of a parallel set of read/write heads, one for each internal state of the box. The output lines 0 and 1 lead to one of the output lines of the core, or to a lazy point that diverts the train into a subroutine that changes the state of that cell on the tape, or to a STOP subroutine that guides the train into a single terminal."

"Let me see," Dee interjected. "For your example, one of the rules is '*State 1, Digit 0:* Change digit, move left, go to state 2.' How does that work?"

"Being in state 1 means that the train enters the cell along line 1, from the side. This state is actually set by the output of the *previous* cell, which directs the train onto line 1 when it exits. In this case, the digit 'written' on the current cell is 0: that is, all the lazy points in the read/write heads are set to 0. So the train comes into the first read/write head, leaves along line 0 and runs into a set of sprung points. These direct it into the 'change digits' subroutine. It runs vertically downward, through all





IN THE CORE* (right), *the train enters from the left, obeys the appropriate Turing machine rule, then exits at the bottom. A layout for a subroutine is shown above: trains enter through lazy points and exit along the same track. Below the subroutine is a read/write head; note the presence of a flip-flop.

of the read/write heads, and flips their states from 0 to 1. So the digit written in the current cell now reads 1, not 0. The train continues back up the vertical track to the left of the heads, exits from the subroutine back onto its original track and then comes out of the core on the output line 2 *left,* which effectively moves the train into the cell to the left, in state 2, as required [*see yellow line on illustration above*]."

"Cute. Suppose we look at the rule '*State 2, Digit 0:* Leave digit, move right, go to state 3.' The train comes in along line 2 and exits the read/write head along line 0, which leads directly to exit 3 *right.* And it never goes anywhere near the subroutine loop, so the state of the cell remains unchanged [*see orange line on illustration above*]."

"Right," Dum said. "And it's equally obvious that the rule '*State 1, Digit 1:* STOP' works properly. Enter along line 1, exit the read/write head along line 1, and you get fed straight into the line that ends at the terminal."

"Exactly. You just set up the lines according to the rules that define the Turing machine."

"Do you realize," Dee asked, "that this shows that the future behavior of a train set can be undecidable?"

"Of course," Dum said. "Turing proved that the halting problem for Turing machines is formally undecidable. You can set up a Turing machine for which there is no way to decide, in advance, whether or not the computation stops."

"Which means that for the corresponding train layout, you can't predict in advance whether the train is ever going to reach the terminal."

"That's quite startling," I remarked. "I've never been terribly bothered by the formal undecidability of theoretical mathematical questions. I mean, who cares? But it's a bit worrying that you can set up a mechanical system—toy train tracks, for heaven's sake—whose workings are totally transparent, but not be able to answer such a simple question as whether the train will ever reach a chosen station."

"Speaking of which—" interrupted Dee.

"It's been an awfully long time since the last stop," Dum said.

I wiped away the moisture fogging up the window. "Hmm," I said. "We've slowed down to a crawl. All I can see is a big square with the digit '1' painted on it. And I swear there's a sign just along the tunnel that reads 'flip-flop 7743A/91.'"

"Dee gets claustrophobic if subway trains stop between stations," Dum whispered.

"They *have* been adding some new connecting lines to the subway network," I offered. "Maybe its connectivity has passed the Turing threshold, and it has achieved artificial intelligence."

"O Mighty Subway," Dee declaimed, her voice rising rapidly in pitch, "we humble humans solicit your omniscient aid in *getting us out of*—" At that moment a connecting door opened, and a uniformed guard stepped into the car.

"Small problem up the line, folks," he smiled. "Nothing to worry about, but we have to slow down for a while." Dee sighed with relief. "Hey, is the little lady all right?"

"Yeah," Dum answered. "She just thought she'd got stuck in an artificially intelligent Turing machine."

"This isn't any touring machine," the guard said indignantly. "This is a personnel commuter, buddy."

At least, I think that's what he said.

FURTHER READING

THE TURING OMNIBUS. A. K. Dewdney. Computer Science Press, 1989.

TRAIN SETS. Adam Chalcraft and Michael Greene in *Eureka,* Vol. 53, pages 5–12; 1994. (To subscribe to this "approximately annual" journal of the University of Cambridge's mathematical society, the Archimedeans, send £10 to the Business Manager, *Eureka,* Arts School, Bene't Street, Cambridge CB3 3PY, England. Internet address: archim@phx.cam.ac.uk)

GREEK WOMEN engaged in all phases of textile work. This illustration, based on a Greek vase of 560 B.C., shows women folding finished cloth, spinning, weaving and weighing out unworked wool.

Unraveling the Past

WOMEN'S WORK: THE FIRST 20,000 YEARS: WOMEN, CLOTH, AND SOCIETY IN EARLY TIMES, by Elizabeth Wayland Barber. W. W. Norton & Company, 1994 ($23).

Two hundred centuries is a long time, and archaeological authors often skimp on evidence of so high an antiquity. Not Professor Barber, whose high expertise supports her unfailing originality of judgment. She shows us the stuff that persuades; her title is apt. The chapters unpack her personal, fascinating and long-awaited synopsis of the textile crafts.

She centers our attention on Europe and its old roots eastward, until womanpower with unstinted talent for putting cloth on our backs at last gave way to steam-turned shafts. Her scholarship is active, wide and deep: few archaeologists these days redraw their own illustrations, even to maps, or cite Homer and Ovid freely in new translations of their own, and fewer still go on to apply the full method: find the data, then "*draw it, count it, map it, chart it,* and if necessary (or possible) *re-create it.*" That activity demonstrably pays.

The source of her evidence for the antiquity of the soft-fiber arts is familiar to most who have read of Paleolithic cultures. It is the well-dated figurines of carved bone, dubbed Gravettian, enigmatic, plump "Venuses" from site after site along the edge of the ice sheet from Spain to Ukraine. One of them, the Venus of Lespugue, is shown with a small skirt of about 12 "long strings hanging down the back from a hip band." The twists in each string are engraved. "Furthermore—a detail I did not notice until I began to make my own drawing from a large…photograph," each string frays out into untwisted fibers. These cannot be sinew or hide, but only twisted threads. From string came cloth and, well before that, fishnets, snares, handles, bindings. "We could call it the String Revolution"; the use of soft plant fibers is of course much older than agriculture. Basketry using hard fibers is even older.

Support is needed, and here it is: a marvelous find of a clay imprint discovered in the celebrated cave of Lascaux, dated 15,000 B.C. The impression records a small length of fiber composed of three two-ply cords, made by people who had the full skills for cordage. More: there are half a dozen string skirts on other figurines of the same culture. Epoch after epoch, the record goes on. In burials of the Danish Bronze Age, just such a skirt, but of wool, was preserved on a young woman, and another one, very brief, "the original miniskirt," with ends weighted by swinging bronze tubes, was found as well. Not at all warm, hardly modest, these costumes were coded symbols, perhaps standing for the readiness to marry.

Your clothing speaks plainly to all you encounter. Its declaratory function is at least as general as its insulating ones; look around you or in the mirror! Later chapters elaborate this conclusion greatly. Barber, adept at these crafts, makes clear that Penelope's famous weaving could not have deceived her suitors for so long were it merely a winding-sheet for a relative's funeral; it was an intricate story-cloth that could plausibly take years for a talented artist to finish. After all, Helen of Troy, so Homer reports, was herself "weaving a great warp,/a purple double-layered cloak, and she was working into it the many struggles/ of the horse-taming Trojans and bronze-clad Achaians."

Over and over we read of textile tools of precious metal, royal gifts to noblewomen; these are more than metaphor, for five or six have been found in Early Bronze Age tombs. So widespread a custom was not likely to have been served by a few eccentric devotees. The palace need for everyday textiles was the work of many servingwomen whose spindles were merely wood. The glittering tools were customary for queens and princesses, who deftly recorded the sacred mythohistory of their clans "with their gold and silver spindles and royal purple wool." (That truly fast colorant was itself precious, gathered drop by drop from rare seashells.) While kings and their sons raised high the funeral mounds and found poets to sing of their heroic deeds, the royal women made an enduring pictorial record on cloth.

"For millennia women have sat together spinning, weaving, and sewing." Why should textiles be so much the women's craft, and was it always so? A sound answer was given by the anthropologist Judith Brown a generation back. We have no example of a community that relied on men for the rearing of children. If productive labor power is not to be lost for many years, a woman must find other work "compatible with simultaneous child watching." Such tasks cannot demand intense concentration; they should be repetitive, easy to resume after an interruption, safe

for children close at hand and easily done at home. Mining, foundry work, plowing will not do. This division of labor is based not on ability but on reliance; males can and often do cook and sew, as females hunt. The question is the practicality of reliance on one group to specialize: for women, hunting is out and textiles in. Adam delved, while Eve as she spun watched and taught little Cain and Abel. Food preparation fits almost equally well: food and clothing were worldwide the core of women's daily work for 20 millennia, perishable artifacts not easily preserved in typical conditions.

This volume would be a reader's pleasure even if its topic were perfectly usual, for Elizabeth Barber is as knowing and perceptive as any archaeologist-author in sight. She has the happy gift of a clear, warm style and a wide interest in image and myth. And her topic is wonderfully fresh—all too much so, for hers is the only recent popular account of the remarkable accomplishments of half our subtle species.

Members of the Club

NUCLEAR WEAPONS DATABOOK, Vol. 5: **BRITISH, FRENCH, AND CHINESE NUCLEAR WEAPONS**, by Robert S. Norris, Andrew S. Burrows and Richard W. Fieldhouse. Westview Press, 1994 ($85; paperbound, $34.95).

Comprehensive if episodic, this volume is based on primary documents of three nations, an archival study of difficulties beside which the frustrations of official Washington seem pale. By turns scrupulously documented and plausibly conjectural, the chronicle reviews the past and present among the Lesser Triads of the Club, their nuclear weapons, production plants and delivery systems.

In America, those most dangerous weapons, nuclear bombs and warheads, are a dwindling species. Although the stock remains fearsome, the crews at the very Pantex plant near Amarillo where these weapons were assembled are now taking them apart instead, five bombs a day—steady, gingerly work for a decade. Similar tasks, less well regulated, occupy Sverdlovsk-45 and a few other numbered secret towns of the former Soviet Union. A tangle of counterpart organizations fills these pages. To match the big plant at Amarillo in west Texas, you might select smaller Burghfield, a few miles from Reading on the upper Thames, or Valduc, a little north of Dijon in the Côte d'Or, or Jiuquan, a Los Alamos–like complex in arid western

Gansu. All four plants named are devoted to the series assembly of nuclear weapons, although the complex manufacture of components—everything from electronics to plastic packaging—and design, testing and eventual transfer to military stocks or to deployment are often far away.

The book allows comparative evolutionary study. Most of the traits converge on the necessities of nuclei (and of governments); some simply record historical links; and a few are no doubt imposed in the uncertain process of analysis. Key nuclear materials for today's nuclear weapons are six in number: normal uranium from mine to metal, its separated light isotope uranium 235, plutonium, tritium, deuterium and lithium 6. These esoteric products are made in similar plants over all three lands; the photographs mainly show the functional ranks of vent-studded factory roofs typical of chemical engineering. Style enters visibly only in France, where some recycled installations preserve the grassy mounds and imposing masonry walls of an earlier world of military engineering.

The explosive devices themselves embody several broad designs that might be called "body plan." Two have been used for fission. First of all came the inefficient gun, assembled by shooting a big bullet of U-235 into a matched target. Obsolescent since Hiroshima, such guns have been stocked as artillery shells and land mines but here appear mostly offstage. (Several such weapons were made and unmade as late as the 1980s by the Republic of South Africa.) Present fission weapons use either U-235 or plutonium (or both) held within a chemical explosive, imploded symmetrically to squeeze some fissile sphere or shell into momentary dense compaction. There is plenty of scope for varying national expertise.

But pure fission weapons cannot easily make light, powerful warheads. Only nuclear fusion enables the bomb designer to meet missile standards; two or three megatons of TNT explosive yield, within one single ton of actual weight, a 200-fold increase over the bulky designs of 1945. Fusion design works within the Three Ideas of Andrei Sakharov: a fission bomb, much boosted in yield by enclosing some tritium within its core; a "layer-cake" design that has not survived in practice; and the two-step device first tested at Eniwetok in 1952. In that ingenious process, first proposed by Edward Teller and Stanislaw Ulam at Los Alamos, a primary fission bomb generates a carefully managed flood of x-rays used to implode the adjoining secondary, a lithium deuteride and uranium charge of more or less unlimited size.

The U.K., China and France have mastered and now deploy deliverable fission, boosted fission and true thermonuclear weapons. They all produce the six essential materials. They all possess powerful fission reactors to make plutonium and tritium as well as plants (mainly gaseous diffusion) to enrich U-235 for bomb and reactor use. They have all tested repeatedly (the dates are listed), none test any more aboveground and probably none will test again anywhere after 1995.

The Chinese have the largest stockpile, some 450 warheads and bombs. The largest single warheads are now also Chinese, mounted on four liquid-fueled missiles targeted on U.S. cities; each yields close to five megatons. Those few long-range deterrent missiles are hidden in silos among many similar decoys, somewhere in China. An up-to-date French warhead, the TN 75, is now under trial at sea. Its characterization here reads like a commercial: miniaturized, hardened (against laser beams), penetrating, safe and "almost invisible" to radar. Its submarine-launched missile carries six warheads to distinct targets up to 3,300 miles away. The U.K. makes its warheads and submarines mainly on its own, but its missiles, like its nuclear tests, have long been American.

All three national "triads" are not as elaborated as the American example. Britain's strategic deterrent relies entirely on the nuclear-powered missile submarines of the Royal Navy. France, too, is steadily reducing its land-based silos. China, never maritime, deploys single-warhead missiles mainly in silos, a fleet of somewhat aging bombers and a truly minimal undersea force.

All these forces arose, of course, from political history. The U.K. was full partner in the World War II origins of nuclear weapons but was kept at arm's length by the U.S. once the war ended. It started independent nuclear armament through a set of secret step-by-step ministerial decisions, completed by the beginning of 1947.

Cooperation between the U.S. and the U.K. returned, especially for missiles, in the early 1960s. In France, postwar work began slowly, then sped up incrementally during the 1950s. The French bomb was long scheduled for 1960.

General Charles de Gaulle had been personally committed to a French bomb as a sign of full independence ever since conversations with French nuclear physicists in wartime Ottawa. A defining anecdote is told here. In 1958 de Gaulle, president of the new Fifth Republic, sought to learn the "disposition of NATO

troops in France." The NATO commander was then General Lauris Norstad, U.S.A.F. military diplomat of the first rank. (It had been Norstad who flew, as 1945 opened, to the big Guam B-29 base, to empower Curtis LeMay to light his 100 jellied gasoline conflagrations in Japanese cities.) President de Gaulle asked Norstad for the locations and targets of U.S. nuclear weapons in France. The tactful NATO general sent out all the entourage before he would answer. Once the two men were alone, Norstad admitted that he was not free to answer the president's questions. " 'General', de Gaulle concluded, 'this is the last time, I am telling you, that a French leader will hear such an answer!' " It was. The first French bomb, a powerful fission implosion, was indeed tested in 1960. France left NATO within a few years. But "the stereotype" of French nuclear independence outlived the reality. Between 1972 and 1985 the French covertly received U.S. "negative guidance" from experts authorized to offer droll hints, winks and nods that led to faster French solid-fuel missile and MIRV design, without giving frankly illegal answers.

The Chinese development started with full Soviet aid, rather as the U.K. began out of its own World War II nuclear partnership. The U.S.S.R. also gave China designs, training, parties of experts, plants and materials, aircraft parts, even Soviet versions of the V-2 itself. (The U.S., too, had tested residual V-2s and drawn leading missile engineers from Peenemünde.) One day in June 1959 all Soviet assistance to China suddenly stopped. Thrown on their own resources, the Chinese made the loss good by enormous effort. It is hard to judge just what had leaked through; it is striking that the Chinese needed only 32 months to pass from their first fission test of 1964, a U-235 implosion, to a full two-stage thermonuclear test. That was less than half the time taken by the U.K., itself the second fastest among all five powers. (The British had minimal U.S. hints about fusion, but there is evidence that they had reinvented the Teller-Ulam configuration as had Sakharov before them.)

The air of change blows through these circumstantial pages. These states are planning reduced forces. The three smaller powers have made over the years no more than 1,000 nuclear warheads each and tested 10-fold less often than have the superpowers. All three of them combined retain roughly a tenth of the unusable arsenals that the superpowers will hold once all present dismantling is complete and all agreements implemented. A new road opens.

Circus Animals

NIGHT AFTER NIGHT, by Diana Starr Cooper. Illustrations by Ivy Starr. Shearwater Books, Island Press, 1994 ($18).

Central within the big blue tent that shelters a rapt audience, the single circus ring is the center of this account, too. The book is no larger than your hand, keenly evocative of an evening's fascination but able to state a lucid view of what humans are: a unique animal, yet one animal species among the animals with which our fate has been entwined ever since we fled the hunting cats and scavenged the lion's leavings.

The circus is the art that most closely explores the relationship of people with animals, says the cofounder of the Big Apple Circus, which is resident in New York City and seasonal visitor among the cities of the Northeast. Cooper glosses his point well: the show reveals people as animals, in cooperation and contrast alike. No animals, no circus.

The word "circus" of course means "circle," the central ring. (Multi-ring circuses are awkward and remain outside present consideration.) The ring is made for horses: its 42-foot diameter was long ago found to be the optimum "to make the back of a galloping horse… hospitable to a person dancing." The horses at the Big Apple are slowly being educated to perform at the world level of haute école by their trainer, herself a lifelong student at the same school. She explains, "Of course we love the horses,/but they are not pets;/they are colleagues./They need me in order to eat,/and I need them in order to eat./They are colleagues/with their own needs and ways/and their own ways of living/and part of that life/we do together."

Circus, a reader infers, is thus neither magic nor wild; it is classical and exquisitely cultivated, unlike the dark and illusory animal mysteries of the shaman. It dates to the time our species worked out a new intimacy, neither hunter nor prey, mutually changing both the animals and ourselves by consciously spending their lives together. Circus is as natural as we are. If you want to see the horses, or Anna May, the artful senior of dancing elephants, or the astonishing subspecies of our lineage, the flyers, the jugglers, the stilt dancers or—get to the Big Apple. For additional reasons, read Diana Cooper's illuminating arguments. She closes in wonder: around that one ring, people and animals are not divided but united. It comes clear that these circles need never end, and "against all opposition, creatures may celebrate the world together."

ESSAY by Devra Lee Davis and Harold P. Freeman

An Ounce of Prevention

Public hopes and presidential promises that cancer could be cured provided much of the cultural meaning and all of the federal funding for the modern war on cancer, launched some two decades ago. The search for a cancer cure, in part, reflected the belief that the disease arises chiefly from discrete external entities that can be attacked and eradicated.

Lately the war on cancer has reinvented itself as the exhilarating quest for defective genes. We read and hear that all cancer is genetic in origin, arising from mutations in the basic building blocks of cells that lead to unregulated growth. Yet only a relatively small portion of most dominant types of cancer is inherited. The key questions remain: What causes the majority of people who have originally inherited a healthy array of genes—some 95 percent of women with breast cancer, for instance—to develop defects that lead them to acquire cancer; and what strategies can be applied to reducing the incidence of disease for all segments of society?

The history of infectious disease in the 19th century supplies a model. Such illness began to decline as a major cause of death largely because of improvements in providing cleaner water, food and air and better housing and working conditions. Reductions in the number of people suffering from these diseases usually stemmed not from treatment but from public health programs that kept these afflictions from occurring in the first place.

Of course, there are more cases of cancer today because there are more older people, and the technology for identifying the disease has advanced. But the rate of all new cases of cancer, excluding that of the lung, has increased about 35 percent since 1950; by one estimate, rates of cases not linked to smoking have tripled in men of the baby-boom generation and grown by a third in women of that generation, compared with their great-grandparents.

In most modern countries, one individual in three will contract some form of cancer, and one in four will die from it. Rates of breast cancer, multiple myeloma and brain cancer have been rising for several decades and are about five times higher in the U.S. than in some Pacific Rim countries and up to 50 times higher than in less developed countries. This remarkable divergence in cancer patterns within and between geographic regions, along with elevated rates in some poor, disadvantaged communities and notably high incidences in about 50 different workforces, implies that a substantial segment of cancer is avoidable or postponable.

At present, the mounting toll of lung cancer provides incontrovertible evidence that smoking remains the single most important avoidable cause of cancer, responsible for about 30 percent of all cancer deaths in industrialized societies. But a prolonged and protracted scientific debate greeted first reports in 1949 from Ernst Wynder that tied smoking to lung cancer. While this debate persisted, millions of people became addicted. Recent reductions in lung cancer in white males in the U.S. and several other developed countries constitute a bona fide public health success, largely attributable to a decline in the number of those who smoke. Unfortunately, smoking-related lung cancer continues to increase at alarming rates in women in the U.S. and in other countries. Also, lung cancer in nonsmokers has reportedly increased in several countries.

A growing body of experimental and human evidence has identified a number of significant environmental risk factors as causes of cancer. They include past diagnostic and therapeutic radiation; diets high in some fats and low in fresh fruits and vegetables; workplace exposures to chemicals, dusts and fumes; pharmaceuticals; sunlight; and heavy alcohol drinking. Long-term, low-level exposures to some environmental contaminants, such as small particulates, chlorination by-products in domestic water and organochlorine residues in animal and fish fat, appear to increase the risk of cancer in human populations, and extensive animal studies indicate a clear risk. Some compounds may function by altering hormones, whereas others may directly affect gene expression.

Meanwhile two decades and $24 billion since the formal launching of the war on cancer, both the war and its warriors are weary. Despite some stunning and gratifying successes in curing the relatively rare cancers of young people, no radically different intervention has been developed for any of the predominant forms of the disease. The poor and uninsured have limited access to early detection and treatment and are often first diagnosed with much higher rates of advanced illness. Even where treatment for relatively rare cancers has been effective, as with testicular cancer, new cases have more than doubled in the past two decades in many countries. Moreover, cancer-cured children and young adults sometimes face a troubling legacy: they have been subject to intensive radiation, surgery or chemical treatments at vulnerable stages of their lives and carry lifelong increases in risk and reductions in function.

This country spends about five times more per patient on chemotherapy than the U.K. does, but survival for most common cancers does not differ. Even when benefit is unexpected, chemotherapy and other cancer treatment have come to be regarded as an entitlement. A decade ago John Cairns of Harvard University pointed out the folly of pouring hundreds of millions of dollars every year into giving a growing number of patients chemotherapy with little proven benefit for the major types of cancer, while doing virtually nothing to protect the population from cigarettes. To this sensible charge, we now wish to add that it is time to turn attention to confirming other avoidable causes of cancer.

No matter how efficient we may become at delivering health care, we must also seek to reduce the need for treatment. An increase in cases of cancer in younger persons in the U.S. and parallel findings in Sweden indicate that we need to identify avoidable causes of cancer in addition to smoking and to develop effective interventions that keep people from developing the disease altogether. If we avert only 20 percent of all cancers each year, we will save more than 200,000 people and their families from this difficult disease and spare the public from the burgeoning costs of treatment and care.

DEVRA LEE DAVIS is senior adviser to the Assistant Secretary for Health at the Department of Health and Human Services. HAROLD P. FREEMAN, director of surgery at the Harlem Hospital Center, is chairman of the President's Cancer Panel.