

SCIENTIFIC AMERICAN

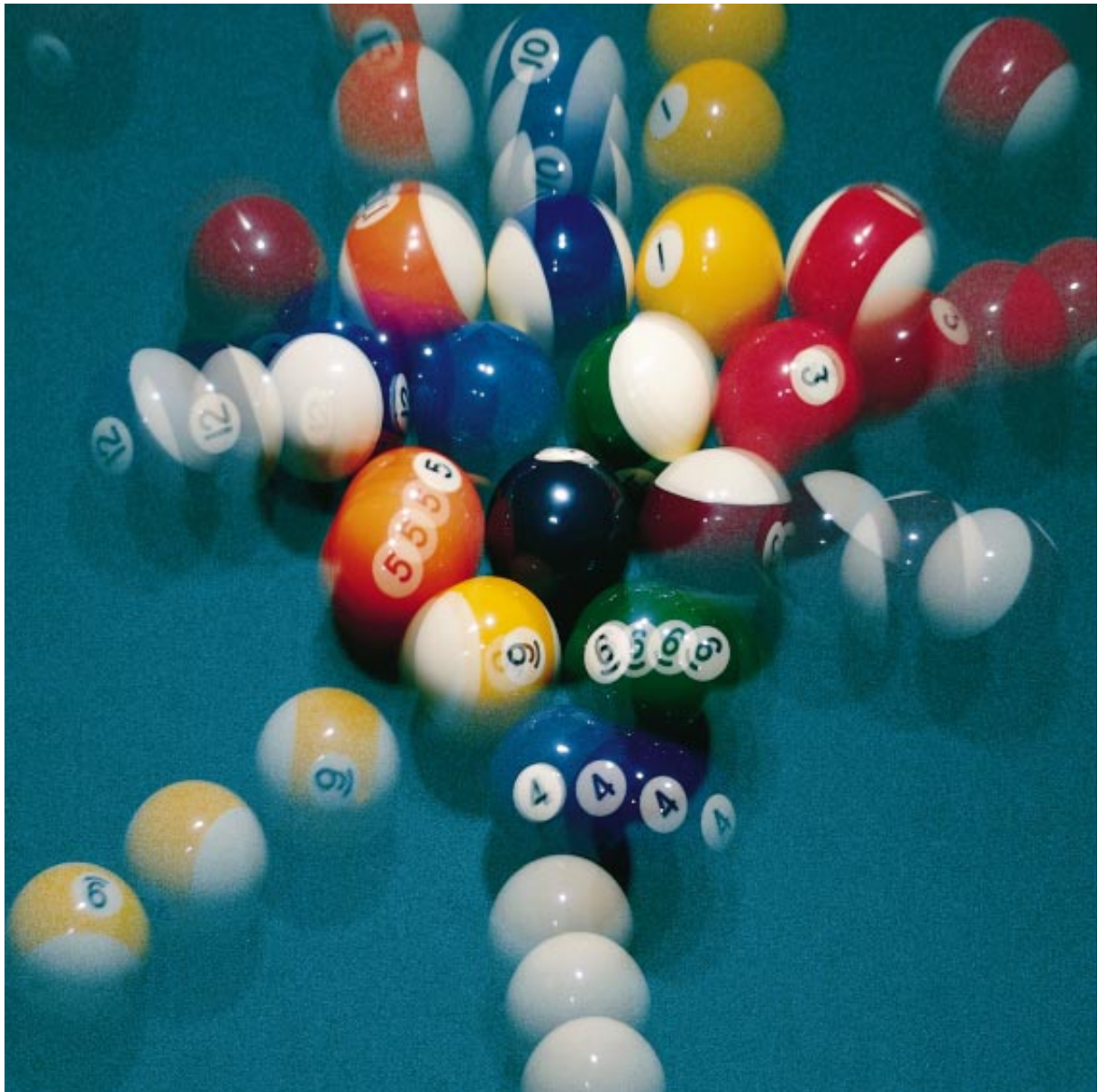
JANUARY 1994

\$3.95

Searching for strange quark matter.

A glimpse at how sex evolved.

The war on cancer: it's being lost.



An even break of the rack creates an intractable problem: calculating the paths the balls will take.

64B



Wetlands

Jon A. Kusler, William J. Mitsch and Joseph S. Larson

Wetlands serve as incubators for aquatic life and shelter higher ground from tides, waves and flooding. But these complex and varied areas are endangered by the demand for real estate, construction sites and cropland. A policy that reconciles society's entrepreneurial endeavors with its need for intact wetlands requires an understanding of these vital ecosystems.

72



The Search for Strange Matter

Henry J. Crawford and Carsten H. Greiner

Protons and neutrons form into atomic nuclei or neutron stars. In between, there is nothing. Nuclear matter does not seem to assemble itself into objects that occupy the range of sizes between these extremes. Yet the laws of physics do in fact permit quarks (the particles from which protons and neutrons are made) to join together to make up objects larger than nuclei but smaller than neutron stars.

78



The Toxins of Cyanobacteria

Wayne W. Carmichael

Cyanobacteria, familiar as a form of pond scum, can be hazardous or beneficial, depending on how one approaches the stuff. As they metabolize, the microscopic single-cell organisms produce proteins and other compounds. These secondary metabolites include potent poisons that can fell cattle and other domestic animals. But they might be co-opted as pharmaceutical agents.

102



Breaking Intractability

Joseph F. Traub and Henryk Woźniakowski

Many important mathematically posed problems in science, engineering and the financial-services industry are computationally intractable. That is, there can never be enough computer time to solve them. But new results indicate some of the problems can be solved if one settles for a solution most, but not all, of the time. The authors also suggest there might be provable limits to scientific knowledge.

108

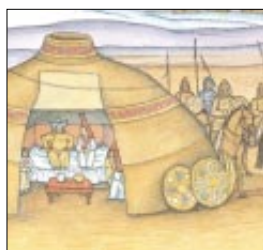


Animal Sexuality

David Crews

Animals have evolved a variety of mechanisms for dictating the division into male and female. In humans and other mammals, chromosomes determine gender. In other species, sex is controlled by temperature or even the social environment. And in a few instances, including a species of lizard, all individuals are female. A new framework for understanding the origin and function of sexuality is suggested.

116



World Linguistic Diversity

Colin Renfrew

Evidence from linguistics, archaeology and genetic studies reveals a pattern of evolution in languages. Today's many tongues seem rooted in a few ancient ones that spread by conquest, the agricultural revolution, the occupation of virgin lands and the dispersal of populations by climatic change.

124

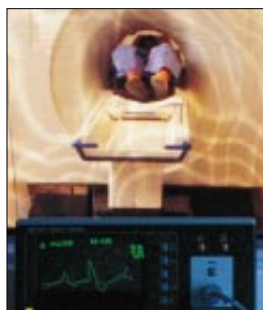


The First Data Networks

Gerard J. Holzmann and Björn Pehrson

Eighteenth-century wireless networks used optical methods to transmit messages. Lines of semaphore stations spanned both revolutionary France and monarchical Sweden. They operated from the late 18th century through the 19th century. Their codes presaged many sophisticated strategies used to transmit data today.

130



TRENDS IN CANCER EPIDEMIOLOGY

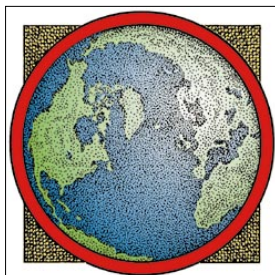
A War Not Won

Tim Beardsley, staff writer

Twenty-five years ago the U.S. declared war on cancer. Since then, billions of dollars have been spent to support tens of thousands of researchers. Surgery, radiation and chemotherapy have been pushed to their limits. Brilliant insights have been gained. And the epidemic sweeps forward. Apart from real progress in controlling some varieties, others remain no more treatable than they were 20 years ago.

DEPARTMENTS

17



Science and the Citizen

Biowar wars.... Turning the NASA battleship.... Dark matter discovered?... Chilling out.... Hot superconductors.... "EQ, phone home".... Dioxin's smoking gun.... Biting the bark.... PROFILE: An all-too-human Albert Einstein.

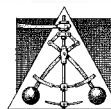
9



Letters to the Editors

Proof lives! Read all about it!... Extraterrestrial inspiration.

12



50 and 100 Years Ago

1894: Why human beings tend to sink rather than swim.

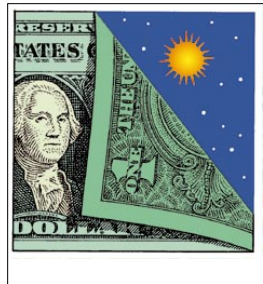
152



Mathematical Recreations

An invitation to a wild evening of knots, links and videotape.

142



Science and Business

Research in recession, the Tokyo touch.... Seeing the light.... Fishy technology.... Here comes biotronics.... Germanium on-line.... THE ANALYTICAL ECONOMIST: Wafty NAFTA models produce future schlock.

155



Book Reviews

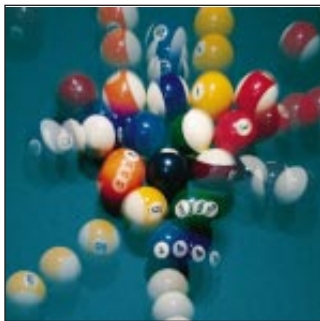
The beauty of bridges.... Rainbows, twilight and stars.... Ancient cells.

159



Essay: George Monbiot

The real tragedy of the commons: a catchphrase reexamined.



THE COVER photograph serves as a metaphor: extreme complexity because of a large number of variables. The 16 caroming and spinning billiard balls render it almost impossible to calculate the dynamics of the break. In fact, solutions to many multivariate problems would require millions of years of supercomputing time. But new theorems indicate that intractable problems can be solved, as long as one settles for what happens most, but not all, of the time (see "Breaking Intractability," by Joseph F. Traub and Henryk Woźniakowski, page 102).

THE ILLUSTRATIONS

Cover photograph by Richard Megna, Fundamental Photographs

Page	Source	Page	Source
64B-65	Stephen Ferry/Matrix	113	Lisa Burnett (<i>top</i>), Pauline I. Yahr (<i>bottom</i>)
66-69	Roberto Osti	114	M. L. East and H. Hofer (<i>left</i>), S. G. Hoffman (<i>right</i>)
70	Cindy Pelescak/South Florida Water Management District	117	Patricia J. Wynne
72-73	Christopher Burke; Quesa- da/Burke Photography (<i>top</i>), Michael Goodman (<i>bottom</i>)	118-119	Dimitry Schidlovsky
74	Edward Bell	120-122	Johnny Johnson
75-77	Michael Goodman	123	Réunion des Musées Nationaux, Paris
79-80	Wayne W. Carmichael	124	Courtesy of Gerard J. Holzmann
81	A. S. Dabholkar, Wright State University (<i>left</i>), Jared Schneidman/JSD (<i>right</i>)	125	Steven Murez/Black Star
82	Jared Schneidman/JSD	126	Guilbert Gates/JSD
84	A. S. Dabholkar (<i>top</i>), Jared Schneidman/JSD (<i>bottom</i>)	127	Gabor Kiss
86	Sushmita Ghosh, University of Illinois (<i>top</i>), Guilbert Gates/JSD (<i>bottom</i>)	128	Courtesy of Gerard J. Holzmann
102-103	Michael Crawford	129	Televerket Tryck & Bild
104-105	Spassimir H. Paskov, Columbia University (<i>top</i>), UPI/ Bettmann (<i>bottom</i>)	131	Berwyn MRI Center/Tony Stone Images
106	Michael Crawford	132	Johnny Johnson
107	National Aeronautics and Space Administration	133	National Cancer Institute
108-109	Patricia J. Wynne	134	Johnny Johnson
110	Patricia J. Wynne (<i>top</i>), Lisa Burnett (<i>bottom</i>)	135-136	Chris Usher/Black Star
111	Gordon Akwera/JSD	137	Johnny Johnson
112	David Crews	138	Jean Louis Atlan/Matrix
		152	Andrew Christie
		153	Michael Goodman
		154	Geometry Center, University of Minnesota

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*; Philip Morrison, *Book Editor*; Corey S. Powell; Philip E. Ross; Ricki L. Rusting; Gary Stix; Paul Wallich; Philip M. Yam

ART: Joan Starwood, *Art Director*; Edward Bell, *Art Director, Graphics Systems*; Jessie Nathans, *Associate Art Director*; 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, *Quality Control*; Tanya DeSilva, *Prepress*; Carol Hansen, *Composition*; Madelyn Keyes, *Systems*; Eric Marquard, *Special Projects*; Leo J. Petruzzì, *Manufacturing & Makeup*; Ad Traffic: Carl Cherebin

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*; William Buchanan, *Manager, Corporate Advertising*; Peter Fisch, Randy James, Elizabeth Ryan, Michelle Larsen, *Director, New Business Development*. 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; Lianne Bloomer. CANADA: Fenn Company, Inc. DALLAS: Griffith Group

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

INTERNATIONAL: EUROPE: Roy Edwards, *International Advertising Manager*, London; Vivienne Davidson, Linda Kaufman, *Intermedia Ltd.*, Paris; Karin Ohff, *Groupe Expansion*, Frankfurt; Barth David Schwartz, *Director, Special Projects*, Amsterdam. SEOUL: Biscorn, Inc. TOKYO: Nikkei International Ltd.; SINGAPORE: Hoo Siew Sai, *Major Media Singapore Pte. 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



LETTERS TO THE EDITORS

Math Abuse

Today's television and movie producers believe violence and death are necessary ingredients for their products. The title and theme of "The Death of Proof," by John Horgan [SCIENTIFIC AMERICAN, October 1993], presumably represent the spread of this belief to *Scientific American*.

The article discussed interesting issues, but it failed to produce the corpse. This is not surprising, since there is no corpse. The true drama of mathematics is more exciting than the melodrama suggested by the title, for this is a golden age for mathematics and for proof. A more appropriate title would have been "The Life of Proof," exemplified by thrilling modern developments, including Andrew Wiles's proof of Fermat's Last Theorem.

The article raised a furor among mathematicians, who, based on the impressions gleaned from its title and spin, became angry at one another for presiding over the death of proof. We were angered at one another—that is, until the dust settled and we compared notes to discover that none of us mathematicians predicts or advocates the demise of proof: we have the common goal of enlivening and enriching proofs.

I need to correct impressions that people have gotten about me from the article. The cover illustrates a scene from the forthcoming video *Outside In*, which presents a proof of a famous theorem due not to me but to Stephen Smale, although the particular proof was devised (many years later) by me. Both *Outside In* and *Not Knot* (in the opening illustration of the article) are explorations of new ways of *communicating* mathematics to a broader public. Contrary to the impression given by the caption "VIDEO PROOF," they are not intended as a *substitute* for logical proofs.

It was suggested in the article that my views sound like those sometimes attributed to Thomas S. Kuhn, to the effect that scientific theories are accepted for social reasons rather than because they are in any objective sense "true." Mathematics is indeed done in a social context, but the social process is not something that makes it *less* objective or true: rather the social processes *enhance* the reliability of mathematics, through important checks and balanc-

es. Mathematics is the most formalizable of sciences, but people are not very good machines, and mathematical truth and reliability come about through the very human processes of people thinking clearly and sharing ideas, criticizing one another and independently checking things out.

WILLIAM P. THURSTON
Director, Mathematical Sciences
Research Institute
Berkeley, Calif.

"The Death of Proof" is certainly thought provoking and very troubling. I agree that computers are causing a revolution in mathematics. I have used them in an experimental way to test hypotheses and even proofs for more than 20 years. I am working on a problem with Matthew Clegg that will eventually involve a calculation using a distributive system of hundreds of workstations. If the outcome of the project confirms the correctness of my hypothesis, then there certainly would be a sense in which the theorem involved would be true. But I have no doubt that a conceptual proof would eventually emerge. This is the crux of the matter to me. Mathematicians should never be satisfied with just "proof"; they should also strive for an elegant proof whose beauty transcends the details that spawned it.

NOLAN R. WALLACH
Department of Mathematics
University of California, San Diego

While I found the article very interesting and well illustrated, I must quibble with the pasta comparison. Helicoids as rotelle? Yes. And by a stretch of the imagination, as fusilli. But helicoids as tortellini? Never!

KAREN WIEDMAN
Altadena, Calif.

Hey, man, thanks a lot for "The Death of Proof." What my buddies down the hall liked best was what you said about how us students don't relate to proofs. We don't. They're real hard, and I don't think we should have to do them, not when you can get the same stuff from those neat color videos. The Grateful Dead likes them, too!

If you guys keep writing neat stories

like this about how math is getting easier and so much cooler, maybe us guys will take some more math courses and maybe even become real mathematicians, 'cause it looks like a real neat job now and not boring like I always thought because of all those numbers and equations and stuff.

Beavis and Butt-head say hi.

BOB MERKIN
Northampton, Mass.

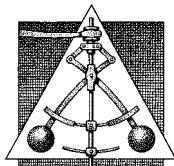
Stars to Wish on

Unlike Richard Wassersug ["Tadpoles from Heaven," "Essay," SCIENTIFIC AMERICAN, October 1993], I believe most people seek their God or ideal not in the heavens but within themselves. Why not take the trillions of dollars that would be spent over several decades to get explorers to Mars and back and use them for studying ourselves—and our nervous systems, in particular? The inner alternative would go a long way toward answering profoundly deep questions, such as how we recognize visual patterns or understand spoken language, as well as "religious" questions concerning free will, evil, compassion and maybe even why we have a religious sense at all.

DAVID G. STORK
Stanford, Calif.

I can vividly recall, as a boy of seven, watching Walter Cronkite follow the launch and recovery of the *Mercury* spacecraft piloted by Col. John Glenn, Jr. I also recall the first manned Gemini flight and the early Apollo flights. I remember the return of detailed images of the surface of the moon and the historic landing of the *Eagle* in the Sea of Tranquility. The risks and accomplishments of NASA throughout the past 25 years have been a constant source of inspiration and admiration. These are the images that helped give me the courage and perseverance necessary to become a productive scientist. I wonder how many of my contemporaries were driven by the same desires and images of future space travel?

TOM NIRIDER
Boeing Defense & Space Group
Seattle, Wash.



JANUARY 1944

“Despite the wide-spread knowledge that forests cannot be indiscriminately logged indefinitely, many pulp-wood producers have been blithely continuing with little or no thought for the future. Result: There is little forestry reserve in the United States today and the vast timberlands of Canada are facing exhaustion. Add to this the other uses for wood that have been developed in recent years—in plastics, explosives, construction work, for examples—and it is obvious that unless something is done, and done vigorously and thoroughly, the paper industry is going to face an even greater crisis after the war than it is facing today.—A. P. Peck, *managing editor.*”

“Two blind spots on the earth’s surface totalling nearly 10,000,000 square miles have been opened up to air travel by one of the most dramatic scientific achievements to come out of the war. Anywhere within 1200 miles of either of Mother Earth’s magnetic poles, magnetic compasses begin to jive and planes enter a shadowy no-man’s-land; this no-man’s-land includes most of Canada. Now, with the gyro flux gate compass, developed by engineers of the Bendix Aviation Corporation, the problem has been solved. The heart of the new compass is three double-wound electromagnets, forming the sides of an equilateral triangle. Different voltages are generated in each magnet, according to the angles at which the compass cuts the lines of force of the earth. Thus the basis of the indication on the compass dial is the combination of the angles and hence of the voltages generated. The resulting current, amplified by vacuum tubes, is stepped up to sufficient power to turn a motor, the shaft of which moves the needle of the dial.”

“The modern trend in the use of chemicals for the control of fire emphasizes prevention rather than fire fighting, says H. L. Miner, manager of the Du Pont Company’s Safety and Fire Protection Division. Mr. Miner notes that paper, cloth, and wood now can be chemically treated to make them incapable of spreading flames. Lumber is chemically being made so fire retardant it is classified on a combustibility scale closer to asbestos than to ordinary wood.”

SCIENTIFIC AMERICAN

JANUARY 1894

“That the continent of Europe is passing through a cold period has been pointed out by M. Flammarion, the French astronomer. During the past six years the mean temperature of Paris has been about two degrees below the normal, and Great Britain, Belgium, Spain, Italy, Austria, and Germany have also been growing cold. The change seems to have been in progress in France for a long time, the growth of the vine having been forced far southward since the thirteenth century; and a similar cooling has been observed as far away as Rio de Janeiro.”

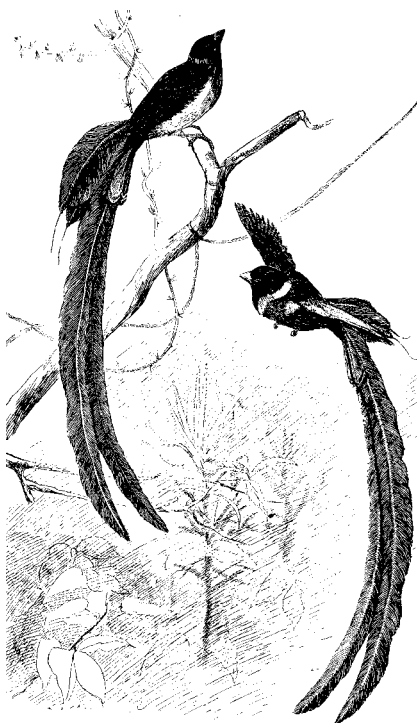
“In a recent article in the *American Journal of Science*, M. Carey Lea gives an interesting account of some of his experiments in which the salts of various substances were subjected to great pressure. The author says: ‘We are justified in concluding that many of the salts of easily reducible metals, especially of silver, mercury, and platinum, undergo reduction by pressure. Such reactions are endothermic, and it there-

fore follows that mechanical force can bring about reactions which require expenditure of energy. The energy is supplied by mechanical force precisely in the same way light, heat, and electricity supply energy in the endothermic changes they bring about.’”

“A writer named Robinson, in *Nineteenth Century*, brings forward a quite plausible explanation for the fact that, while most of the animal creation appear to swim by intuition, man is almost alone in requiring previous training to keep his head above water. He says it is due to our descent from races who were cave and rock dwellers and rock and tree climbers. Robinson suggests that the hereditary instinct of man is unfortunately to *climb* out of danger. Hence, unless he has a natatory education, he throws his arms at once above his head, thus increasing the weight upon the latter, which of course, goes then under water.”

“Mlle. Klumpke, who has just gained the degree of Doctor in Mathematical Sciences at the Sorbonne, is the first lady who has obtained that distinction. The following is a translation of the complimentary terms in which M. Darboux addressed the gifted authoress in granting her the degree: ‘The great names of Galileo, Huyghens, Cassini, and Laplace are connected with the history of each of the great advances in the attractive but difficult theory of the rings of Saturn. Your work is not a slight contribution to the subject. The Faculty has unanimously decided to declare you worthy of the grade of Doctor.’”

“Through the kindness of Mr. W. Stoffregn, importer of birds, we are enabled to give a representation of the beautiful widah bird of paradise. It is an inhabitant of Western Africa. The male bird in his full dress is a deep black on the wings, tail, and back, with a collar of bright yellow. The head and throat are also black, the breast being a rich reddish-brown. The bird has been commonly called the widow bird on account of its dark color and long train, as well as in consequence of its evidently disconsolate state when the beautiful tail feathers have fallen off after the breeding season. The widah bird measures between five and six inches, exclusive of the elongated tail feathers.”



The widah bird of paradise



Joe Btfsplk

NASA's big-science projects find themselves on a rocky course

For his Li'l Abner cartoons, Al Capp dreamed up a character named Joe Btfsplk—a man so unlucky that a tiny raincloud followed him wherever he went. Although the artist and the original comic strip are gone, Joe apparently has a new job: patron saint of the National Aeronautics and Space Administration. And he's been working overtime. In the past few months, the agency has experienced a seemingly endless string of bad fortune, including the mysterious, mission-destroying loss of contact with the *Mars Observer*. Even the *Galileo* spacecraft's successful encounter with the asteroid Ida last August was compromised by an incurable antenna problem that has significantly reduced the probe's ability to relay information back to the earth.

Some setbacks are inevitable in space science; no rocket is perfectly reliable, no instrument foolproof. But NASA's recent problems arouse particular disappointment and frustration because they involve big-science projects whose failures carry an especially heavy cost to the taxpayers and to the scientists involved. Despite the "cheaper, faster, better" philosophy espoused by NASA's current administrator, Daniel S. Goldin, unwieldy scientific behemoths remain alive if not always well at the agency.

The *Mars Observer* stands as a telling example of how hard the task of turning the NASA battleship can be. More than a decade ago the vehicle was proposed as the first of a new generation of economic, efficient "Observer-class" spacecraft. They were to embody a common design and be furnished with low-cost, off-the-shelf technology. If that description sounds familiar, it should. NASA has set similar goals for its proposed "Discovery-class" missions, the first of which, ironically, will go to Mars. "Discovery is where the Observer missions were 10 years ago," reflects Larry W. Esposito of the University of Colorado, who is currently drawing up plans for a possible Discovery mission to Venus.

The Observer program never won over Congress or the Office of Management and Budget, however. So the *Mars*



NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

U.S.-RUSSIAN SPACE STATION, shown in this computer-generated mockup, hints at a new, international spirit that may help revive NASA.

Observer became a one-of-a-kind orphan. The cost savings associated with building multiple spacecraft vanished, and the *Mars Observer* grew more complicated and expensive as space scientists and NASA officials tried to expand its capabilities as much as possible. When the space shuttle *Challenger* exploded in 1986, the *Mars Observer* encountered extensive delays that drove its price even higher.

Even before reaching Mars, the Observer project had consumed roughly \$850 million. For that money, NASA put together a sophisticated suite of instruments designed to convey information on the geology, mineralogy and climate of Mars. It would have been the first U.S. mission to the Red Planet since *Viking* in 1976. Unfortunately, the *Mars Observer* stopped communicating just before it reached its destination. As John Pike of the Federation of American Scientists points out, the loss of the *Mars Observer* underscores NASA's need for "a selection process that does not encourage everyone in the scientific community to put all their eggs in one basket."

Indeed, NASA's follow-up strategy for exploring Mars already envisions cheaper and more diversified missions. In 1996 NASA hopes to launch a technology test bed for the *Mars Environmental Survey (MESUR)*, which would form part of a network of as many as a dozen low-cost scientific stations scattered across the surface of Mars. *MESUR* may establish a more international flavor at NASA. At a meeting last May in Wiesbaden, Germany, representatives of the world's major space programs, including NASA, the European Space Agency and the Russian Space Agency, met to coordinate their plans for exploring Mars. Louis Friedman, executive director of the Planetary Society, heartily endorses NASA's newfound cooperative spirit, although he worries that efforts to involve international partners in *MESUR* "haven't gone far enough."

For the moment, Congress seems to agree that NASA is on a promising trajectory; the tentative 1994 appropriations bill for the agency significantly increases funds both for *MESUR* and for the second Discovery mission, the *Near Earth Asteroid Rendezvous*. "I'm very

optimistic," Esposito says. "It shows that NASA and Congress are committed to flying faster, cheaper missions."

While Goldin attempts to nudge NASA toward more small, high-tech ventures, he must also make the best of several troubled big-science projects already under way. "It's ironic, but Goldin's success is linked to having to fix the mistakes of the past," notes John M. Logsdon, a space policy analyst at George Washington University. NASA has already devised fixes for the nearsighted *Hubble Space Telescope*, and *Galileo* continues to transmit valuable scientific results despite its faulty antenna.

In response to congressional pressure, NASA has also placed several upcoming missions on budgetary diets. The agency has pared back both the Cassini mission to Saturn and the ambitious fleet of satellites that will make up the *Earth Observing System*. The *Advanced X-ray Astrophysics Facility*, a satellite observatory that would complement *Hubble* and the *Compton Gam-*

ma Ray Observatory, has been split into two smaller instruments, only one of which is on track to receive congressional funding. Pike dryly remarks that "so far 'cheaper, faster, better' has turned out to mean 'less.'"

Not surprisingly, the space station—NASA's porkiest project—is also in dire political trouble. The station is already years behind schedule and billions of dollars over the budget envisioned by President Ronald Reagan 10 years ago. Last summer a measure in the House of Representatives to kill the station failed by just one vote. Yet although Congress subsequently terminated the Superconducting Super Collider, the station soldiers on.

The space station's new lease on life is financed by the growing detente between the U.S. and Russia. Last August, Vice President Al Gore and Prime Minister Viktor S. Chernomyrdin signed an accord promising cooperation between the two nations' space programs. Goldin recently outlined a three-stage plan

to combine the revamped space station *Alpha* with the Russian station *Mir* by 2001, two years earlier than the current schedule for *Alpha* alone. Goldin claims such an arrangement could save up to \$3.5 billion. Meanwhile he is drastically cutting the size of the space station management team.

So, ironic though it may seem, the battered and bloated space station might yet be the vehicle that carries NASA into a future characterized by the efficiencies that should accompany international cooperation. The remodeled space station, Friedman says, could serve as the core of an internationally conscious NASA that will move away from massive, autarkic projects such as the *Mars Observer*. To accomplish such a change, NASA will need, in Pike's words, "significant restructuring": stronger long-range planning and more efficient management (and, of course, a small bout of good luck). Time will tell whether Goldin's team at NASA can exorcise Joe Btfsplk. —Corey S. Powell

Getting a New Rise out of Superconductors

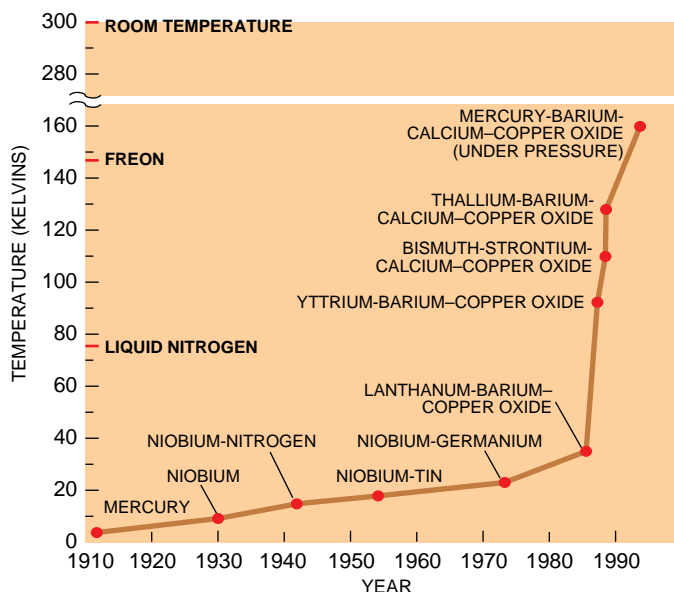
Avoiding pressure is usually good advice—but not for scientists trying to get ceramics to become superconducting at higher temperatures. Indeed, putting the squeeze on mercury-barium-calcium-copper oxide, a new family of ceramic superconductor discovered last year, has boosted its transition temperature to record levels. "We now have a new set of results of 164 kelvins at 300 kilobars [about 300,000 atmospheres]," says Paul C. W. Chu of the University of Houston.

The as yet unpublished result comes on the heels of two other high-pressure reports, one by Chu and the other by Manuel Nuñez-Regueiro of the CNRS in Grenoble and their colleagues. The groups found that the mercury compound, called 1223 (for the ratio of the compound's first four constituent elements), becomes superconducting above 153 kelvins at 150,000 atmospheres and 157 kelvins at 200,000 atmospheres. Those critical temperatures mean the com-

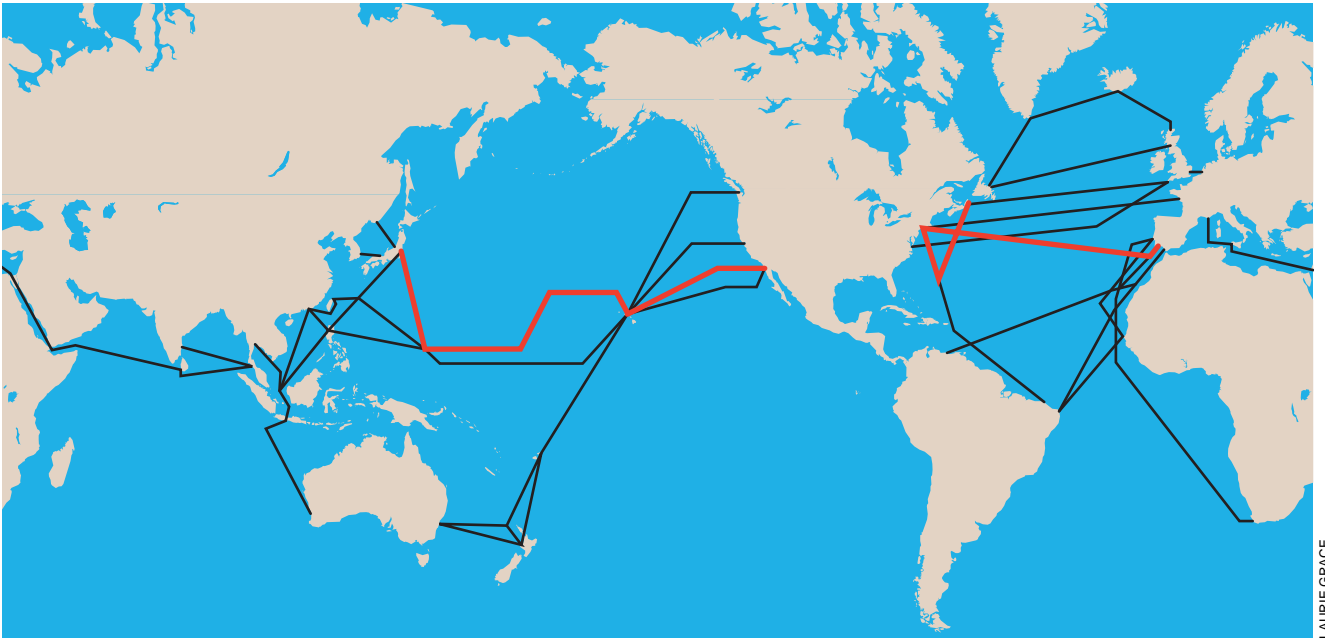
pounds could be cooled with the common (but environmentally hostile) coolant freon. The pressure, achieved by placing a sample in a vise, apparently moves the layers of copper oxide in the material closer together. For some unknown reason, the proximity enables the electrons to flow more freely. The investigators hope to sidestep the high pressures, which render the results impractical for ap-

plications, with a chemical substitution. By replacing one of the elements with a smaller one, they would lessen the distance between copper oxide layers. In fact, Chu and his colleagues used such a strategy to discover the superconductor yttrium-barium-copper oxide in 1987.

The surging competition is reminiscent of the early days of high-temperature superconductivity, when records seemed to fall every few months and unconfirmed reports hinted at superconducting transitions at room temperature. Although the new mercury oxides have reinvigorated the chase, physicists will not be dumping their supply of cryogen just yet. The mercury compounds do not seem to be able to go much higher. "At this moment, the empirical data suggest we can go to 180 kelvins," Chu says in a somewhat disappointed tone. But the 180-degree view still shows just how far critical temperatures have come since superconductivity was discovered in 1911. —Philip Yam



CRITICAL TEMPERATURES remained below 23 kelvins until the discovery of the copper oxides in the late 1980s.



LAURIE GRACE

TELECOMMUNICATIONS CABLES stretch across thousands of kilometers of ocean where geophysical data are not currently available. This map shows the coaxial cables that are being joined or replaced by fiber-optic lines; those shown in red may soon assume a second, scientific life as part of an undersea seismic and oceanographic network.

“EQ, Phone Home”

Undersea telephone cables could serve as seismic detectors

Connectivity is the way of the 1990s, and earth scientists are getting in on the act. They have a new mission for the transoceanic telephone wires that AT&T and other long-distance telephone companies are rapidly replacing with fiber-optic cables. Over the past few years, a number of earth scientists, including Charles Helsley of the University of Hawaii, have proposed that the obsolescent cables could provide the infrastructure for a network of instruments that would monitor earthquakes, ocean currents and other aspects of the deep-ocean environment. “There’s a lot of copper that crosses the oceans,” Helsley comments. “It’s just a millstone around the company’s neck, but it could be very valuable from the scientific point of view.”

Telephone cables offer a way to get power into and information out of devices in such remote locations as the Indian Ocean and the southern Pacific. They can also deliver accurate timings of seismic events in out-of-the-way places, notes Rhett Butler of the Incorporated Research Institutions for Seismology (IRIS). Right now seismometer coverage is “just about zero in the oceans except for a few islands,” Helsley says.

Many of these cables cover areas of great scientific interest. Alan Chave of the Woods Hole Oceanographic Institution points to Transatlantic-5, a cable that passes through the Gulf Stream and crosses the Mid-Atlantic Ridge. Even the cables that are less attractively located could be pulled up and redeployed in more interesting places.

The dream of assembling a suboceanic seismic network moved sharply toward reality four years ago, when the University of Tokyo and IRIS assumed control of a stretch of Trans-Pacific Cable-1, which extends from Guam to Japan. Plans called for splicing three seafloor observatories into the cable. Completion of that project awaits solution of funding problems in Japan. AT&T has been generous about donating old cables, but hauling them up from the seafloor and attaching instrumentation are quite costly—about \$1 million a splice, estimates Charles S. McCreery, also at the University of Hawaii.

McCreery and various colleagues of his are looking at a cheaper way to get on-line. McCreery is investigating devices that would attach to the telephone cables without penetrating them and would magnetically induce an electrical signal. Such an approach could be done at “an order of magnitude less cost,” he suggests. Time is of the essence in building an undersea network. “Cable systems are being retired from service faster than the scientific community can mobilize funding to acquire the systems for science,” according to

a recent IRIS report. “The first priority is to save the shore equipment,” Butler says. Two transatlantic cables have already been torn out and their shore equipment decommissioned.

Fortunately, some scientific work on abandoned cables needs only basic instrumentation—and hence very little money. Natural electric currents exist in the oceans because of fluctuations in the earth’s magnetic field, the interaction of that field with oceanic circulation, and changes taking place deep within the earth’s metallic core. Monitoring the electromagnetic phenomena necessitates little more than attaching an exceedingly sensitive voltmeter to a telephone cable and watching what happens over periods ranging from days to years.

Such information will help researchers map the electrical conductivity of the outer layers of the earth and should yield sharper understanding of large-scale ocean circulation. Preliminary studies conducted on the Hawaii-1 cable in the eastern Pacific look promising. Chave recently received a two-year grant from the National Science Foundation to attach instruments to a leg of Trans-Pacific Cable-1.

For now, funding for ocean-bottom observatories is “modest, very modest,” in Butler’s words, so researchers are scaling their plans accordingly. As Helsley jokingly puts it, he and his colleagues just want “a telephone booth on the seafloor we can hook a modem onto.”

—Corey S. Powell

A Dark Matter

Astronomers may be closing in on the invisible cosmic majority

Anybody who ever doubted that nature has a perverse sense of humor should consider the plight of the astronomers trying to map out the structure of the cosmos. Most of the mass of the universe seems to exist as some form of “dark matter” that is invisible through any kind of telescope. Studies of how galaxies rotate and move about one another indicate that they are enveloped in halos of such material. But researchers do not know what dark matter is made of. They have considered everything from undiscovered subatomic particles to snowballs floating in space.

Now at last they have a clue. Three teams have made observations hinting that at least some of the dark matter surrounding our galaxy consists of diminutive relatives of the sun: faint, low-mass stars and brown dwarfs, objects larger than planets but still too small to shine like stars. Kim Griest of the University of California at San Diego has collectively dubbed such objects MACHOs (massive compact halo objects)—a riposte to his particle physicist colleagues who propose that dark matter is composed of WIMPs (weakly interacting massive particles).

The key question that has daunted researchers attempting to learn about dark matter is, How can one identify something that cannot be seen? In 1986 Bodhan Paczyński of Princeton University realized that astronomers could, in principle, perceive the gravitational tug produced by MACHOs even though the objects themselves are nearly undetectable. Einstein’s theory of relativity states that gravity can bend light. If a MACHO were to pass between the earth

and a more distant star, its gravitational field would act as a magnifying lens, bending and focusing light from the background star. Because of that effect, the background star would appear brighter than normal. As the MACHO continued on its path, it would move out of alignment, and the star would return to its usual brightness.

Paczyński realized that searching for such an event—known as gravitational microlensing—would require monitoring the exact brightnesses of huge numbers of stars over an extended duration. “In 1986 it was science fiction—the technology wasn’t there to monitor a million stars,” Paczyński recalls.

Since then, improved digital light detectors and high-speed computers have swiftly transformed fiction into a practical reality. By 1993 at least three sets of investigators (a U.S.-Australian team led by Charles Alcock of Lawrence Livermore National Laboratory, a U.S.-Polish group led by Paczyński and a French collaboration headed by Michel Spiro of the Saclay Research Center in France) had begun a determined hunt for the blips of light that might settle the dark matter question. Last fall all three teams reported tentative sightings of the microlensing phenomenon—a rapid-fire succession of results that Paczyński refers to as “stimulated emission.”

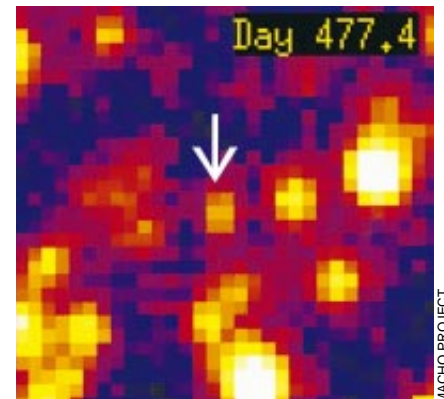
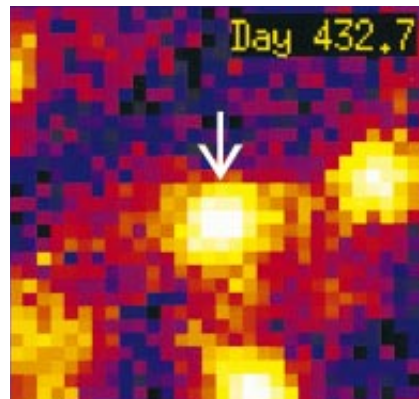
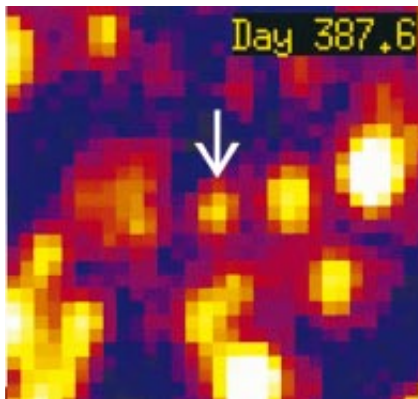
Griest, who participates in Alcock’s group, recounts that he and his colleagues had been monitoring 1.8 million stars in the Large Magellanic Cloud, one of the Milky Way’s satellite galaxies, for nearly a year without detecting anything unusual. “We were ready to put upper limits on the amount of MACHO dark matter when out popped a good event,” he reports. As the news spread through the collaboration, rumors began to circulate that the French team had just recorded an event of its own. The two groups ended up making simultaneous announcements. Shortly

thereafter Paczyński and his co-workers announced a third, similar event seen toward the center of our galaxy.

All the observed events display one of the most telling characteristics of microlensing: a slow brightening followed by a perfectly symmetrical dimming. No known kind of variable star or other astronomical object would show such a pattern. Moreover, the French and U.S.-Australian groups can demonstrate that the stars did not change color during the event—a trait expected of microlensing but one not shared by known variable stars.

So have astronomers finally solved the riddle of the dark matter? Well, not exactly. First of all, the researchers could be looking at a new kind of variable star. Second, the data are impressive but by no means perfect. Griest points to a strange-looking data point in his light curve that “still makes me nervous.” And the identity of the microlensing objects remains ambiguous. Based on the duration of the detected events, the three groups calculate that they have probably recorded bodies much less massive than the sun. But such estimates contain considerable uncertainty; the objects detected so far could actually be solar-mass stars, which emit too much light to make up a substantial part of the dark halo of the Milky Way.

The researchers are racing to analyze more data so they can establish useful statistics on the total amount of matter tied up in dark, low-mass MACHOs. “We’re cranking really hard,” Griest replies, more than once, when asked about his group’s progress. That eagerness to uncover a previously undetected component of the universe—one that may outweigh all the visible stars in the night sky—is easy to understand. As Griest reflects, if his results pan out, “we’re starting a whole new field of astronomy.” —Corey S. Powell



STELLAR BRIGHTENING (seen in the center of these digital images) is thought to result from the gravitational pull of an unseen body—possibly the long-sought “dark matter”—that passed between the earth and a more distant star.

Biowarfare Wars

Critics ask whether the army can manage the program

Over the past decade, the U.S. has spent more than \$600 million trying to anticipate and develop defenses against an attack involving biological weapons. The primary justification of the so-called biological defense program has always been the Soviet Union, which was alleged by past U.S. administrations to have a vigorous offensive program—in violation of the Biological Weapons Convention. Now that the cold war is over, some arms-control advocates are contending that the U.S. should curtail its research into such weaponry and concentrate on stemming proliferation through international agreements.

Yet the need for defenses against biological weapons—such as detectors, protective clothing and vaccines—is more compelling than ever, according to military officials. Advances in biotechnology, they assert, have made biological weapons an increasingly attractive alternative to countries whose resources would not be sufficient to develop a nuclear arsenal. The Pentagon claims that as many as 25 nations, including such avowed enemies of the U.S. as North Korea, Iran and Iraq, are now developing biological weapons or have already done so. Billy Richardson, who as deputy assistant secretary of defense for chemical matters oversees both chemical and biological defense research, has testified before Congress that “biological warfare defense has gained unparalleled interest and support” within the Pentagon and has been designated a “priority requirement” by senior military officials.

The Department of Defense has requested some \$60 million for its research program for 1994, up from \$50 million in 1992. The army is seeking funds for a new vaccine-testing facility at Fort Detrick, Md., which has been the headquarters for biowarfare research since World War II. Moreover, last June the army announced its intention to construct a laboratory for testing pathogens at the Dugway Proving Ground in Utah. In the mid-1980s opposition from grass-roots groups and such prominent Utah politicians as Senator Orrin Hatch blocked plans to build a facility at Dugway for research on the most dangerous agents that might be developed, notably genetically altered pathogens for which there is no cure. The army now intends to erect a facility that has less rigorous containment

features but is still qualified to handle such deadly agents as anthrax, botulin toxins and encephalomyelitis viruses.

Has the money allocated thus far to the biological defense program been well spent? This question has been raised not by the military's traditional critics but by the General Accounting Office. One GAO report found that at the beginning of the Gulf War the U.S. Army's stockpiles of vaccines for anthrax and botulism, which were thought to make up the bulk of Iraq's biological arsenal, fell far short of what was needed to protect U.S. troops. In 1990 the GAO concluded that at least 20 percent—possibly as much as 40 percent—of the army's biological weapons budget was not directed at diseases or toxins identified as threats by the military's own intelligence. In fact, the GAO found that the army “may unnecessarily duplicate medical research” on vaccines already being done at the National Institutes of Health and the Centers for Disease Control.

Pentagon officials respond that no civilian agency can address military needs and questions. They also argue that the shortcomings exposed by the Gulf War show that the program needs more support, not less. Yet critics of the biological defense program have urged that research involving vaccines and other medical applications requiring the handling of live pathogens be placed under a civilian agency. Last June, Congress took a step toward that goal. Lawmakers have required the Department of Health and Human Services to study the “appropriateness and impact of the National Institutes of Health assuming responsibility for the conduct of all Federal research, development, testing and evaluation functions relating to medical countermeasures against biowarfare threat agents.” The health secretary's report is due next June.

By at least partially demilitarizing its program and thus making it more open to scrutiny, might the U.S. aid international arms-control efforts? According to Susan Wright of the University of Michigan, a political scientist and an authority on biological weapons, the answer is affirmative. “Whatever the U.S. does is going to provoke attention and be copied to some extent,” she remarks. For several years, arms-control groups have been urging the adoption of verification provisions to enhance the Biological Weapons Convention, which prohibits the manufacture and use of biological weapons as well as offensive research. The convention has been signed by more than 120 countries, including the U.S., since 1972.

In 1991 signers of the convention established committees of experts to study verification. The experts presented their reports at a United Nations forum last fall, and members are expected to begin formal negotiations of verification provisions sometime this year. Such provisions could call for both routine and unscheduled inspections of industrial and governmental biotechnology facilities as well as requiring detailed annual reporting on dual-use activities. The Reagan and Bush administrations opposed such measures, contending that they would be ineffective and would lead to disclosures of proprietary information.

“Is perfect verification possible?” asks Barbara H. Rosenberg of the State University of New York at Purchase, who heads the chemical and biological weapons verification project of the Federation of American Scientists. “Everyone agrees it isn't, especially for biological weapons that involve dual-use technologies. But it's aimed at providing more openness.” To encourage developing countries to submit to intrusive verification, she adds, advanced nations might have to help them acquire biotechnology by relaxing export controls. “All the developing countries are interested, but nothing has happened yet,” she says.

The Federation of American Scientists and the World Health Organization are also seeking to make the verification regime part of a broader effort to monitor and respond rapidly to the outbreak of diseases, whether caused deliberately or naturally. The two organizations sponsored a meeting in Geneva last September to consider the plan, called the Program on Monitoring Emerging Diseases.

But arms control alone is not enough to protect U.S. troops, according to a member of a congressional committee with oversight of the biological defense program. She rejects Wright's contention that the U.S., by cutting back on or demilitarizing its biowarfare research, might discourage other countries from acquiring biological weapons. Such an act “won't stop North Korea or Iraq or Iran” from developing such weapons, she asserts.

The Clinton administration has yet to set forth an explicit policy on its own biological defense program or on arms-control efforts. An administration source suggests that although the White House may support more intrusive arms-control measures, it is unlikely to curtail or demilitarize its own effort. “My own view,” the official notes, “is there is a real need for a strong biological defense program.” —*John Horgan*

Chiller Thriller

Workers achieve temperatures below absolute zero

Research in physics has reached a new low. Scientists at the Helsinki University of Technology have measured picokelvin (trillionths of a degree) temperatures just above, and even below, absolute zero in metal-

lic rhodium. These temperatures are much lower than any previously recorded. When asked what the feat means, Pertti Hakonen, leader of the Finnish team, plunges into a review of the dynamics that describe temperature. By definition, temperature measures the energy, or the amount of disorder, in a system. A system having absolute zero temperature would be unquestionably free from all atomic motion. As a result, the system would hold no energy

and no entropy. The electrons in the lattice of a crystal would, for example, be utterly still. The spins in an array of atomic nuclei might all point in the same direction (think of a clutch of tiny planets spinning in space).

But there is a catch. The third law of thermodynamics states that such a condition could not happen. The particles that make up all matter must vibrate, at least a little, all the time. Following ordinary logic, then, it would

Something to Chew on

By chewing on the bark of a white willow tree, Edmund Stone, an 18th-century Anglican clergyman, discovered the analgesic merits of salicylic acid, the active ingredient in aspirin. No one, no matter how grateful for pain relief, has yet fathomed why Stone was gnawing on willow bark. But a possible reason why the willow and other plants produce this versatile compound has been discovered. A team from the Agricultural Biotechnology Research Unit at Ciba-Geigy has shown that the accumulation of salicylic acid in plant tissue after an infection is essential for prompting a crucial immune response, called systemic acquired resistance (SAR).

The two main defenses a plant inherits to fight disease are known as vertical resistance and horizontal resistance. Vertical resistance acts against individual agents of disease. Horizontal resistance, a category to which SAR belongs, is mounted against a wide array of related plant

pathogens. It works by stalling fungal, bacterial or viral proliferation and activity. Because horizontal resistance protects against many kinds of plant pathogens, the ability to mobilize SAR in the absence of an actual infection could bolster a plant's ability to ward off disease. "One of our goals is to develop chemicals to spray on plants that will actually trigger a plant to be healthy," says John Ryals, the project's research director.

Systemic acquired resistance appears to be involved in the control of the expression of a set of genes that encode for specific proteins. Some of these proteins act like antibiotics when tested against plant pathogens in vitro. These proteins may help keep a plant healthy when exposed to disease. An external application of salicylic acid to tobacco leaves causes SAR to develop quickly as though a pathogen were present.

Work by the Ciba-Geigy researchers reported in a recent issue of *Science* confirms that the onset of SAR is related to a plant's salicylic acid levels. Ryals and his colleagues wrote that by blocking the buildup of salicylic acid in infected tobacco plants, they had weakened the plants' ability to resist infection. Specifically, they prevented the accretion of salicylic acid in tobacco plants by inserting a gene for producing salicylate hydroxylase, an enzyme that breaks down salicylic acid.

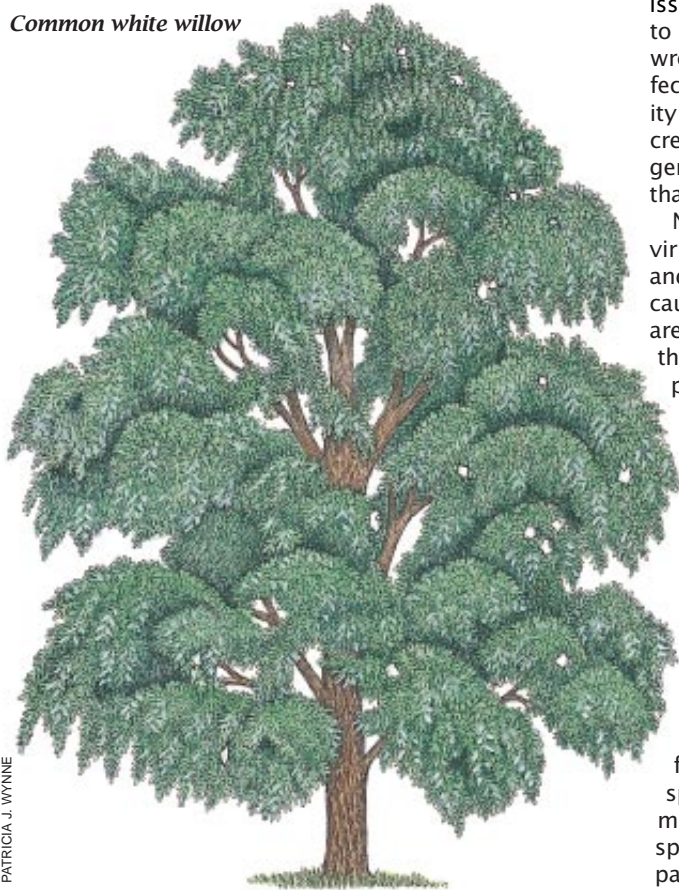
Next the researchers inoculated the tobacco mosaic virus (TMV) into three lower leaves of the altered plants and of the unaltered, control-group plants; the disease causes splotches of dark-green blisters and dulled yellow areas. Seven days after the lesions appeared, members of the Ciba-Geigy laboratory harvested the leaves and compared them. Leaves from the control group showed much less damage. Those plants had also accumulated an expected 185-fold increase in salicylic acid after the infection. The specimens in which the salicylate hydroxylase gene had been implanted showed only minor increases in salicylic acid.

The workers then exposed the upper leaves of the plants infected with TMV to a second dose of the virus. Five days later the leaves that were low on salicylic acid had the largest lesions. This result confirms the harbinger role the chemical plays in this form of plant immunity.

Although these data demonstrate that salicylic acid must be present for the development of SAR, other factors are known to be involved in controlling the response. When investigators have deciphered the entire mechanism controlling SAR, the secrets revealed could spare plants from physical ills and farmers from financial pain as well.

—Kristin Leutwyler

Common white willow



PATRICIA J. WYNE

seem impossible to attain temperatures below zero. The secret of the Finnish group's success, Hakonen notes, is that negative temperatures are in fact not colder than absolute zero.

In their laboratory, Hakonen and his colleagues measure nuclear spin temperatures. First, they place a substance in an external magnetic field, so that the nuclei will spin parallel to the external force in numbers proportional to the field's strength. When the majority of the nuclei spin in the same direction, the sample registers a low, positive nuclear spin temperature. This high degree of parallel, or ferromagnetic, order coincides with the lowest energy level and least entropy available to the system.

Next, the physicists quickly (within the span of a millisecond) flip the direction of the applied magnetic force. Most of the nuclei then spin in opposition to the external field in high-energy orientations. The process is adiabatic, meaning the entropy remains unchanged. The resulting spin distribution is the inverse of that associated with positive nuclear spin temperatures. Hence, it is assigned a negative value. "The main difference is that at a negative temperature, the system tries to maximize its energy," Hakonen explains.

In a sense, negative temperatures can be considered hotter than infinite temperatures. An infinite nuclear spin temperature correlates with an even distribution of possible spin alignments: just as many nuclei assume high-energy orientations as do low-energy ones. The arrangement represents maximum entropy, or chaos, within a substance.

Heating such a material forces growing numbers of nuclei to spin in opposition to the external field in order to absorb the additional energy. The probability of any given nucleus assuming a high-energy spin orientation increases, and so overall entropy in the system decreases.

By coaxing substances to low temperatures very near absolute zero, physicists have hoped to observe the weak magnetic interactions that transpire between neighboring nuclei. This complicated pattern governs how each individual spin affects the next, in a dominolike fashion throughout the material.

Hakonen and his colleagues, who reported their work in *Physical Review Letters*, detect the spin orientations of rhodium by recording nuclear magnetic resonance spectra with a SQUID magnetometer. At the moment, they are preparing experiments for cooling platinum. Frosty femtokelvin (quadrillionths of a degree) temperatures may yet be within reach—particularly during the long Finnish winter. —*Kristin Leutwyler*

Dioxin Indictment

A growing body of research links the compound to cancer

Dioxin has always seemed a paradoxical pollutant. In laboratory animals, it is clearly a potent carcinogen; in humans, its link to cancer has been tenuous. But a recently published study of people exposed to the toxin presents compelling evidence that dioxin has carcinogenic effects in the human species as well.

Since 1976, when an industrial accident spewed dioxin into the air near the Italian town of Seveso, scientists have monitored the health of about 2,000 families there. Several years ago the researchers documented increases in cardiovascular disease and suggestive increases in certain cancers.

More current work by the same group has strengthened the evidence for dioxin as a carcinogen in humans. Writing in *Epidemiology*, Pier Alberto Bertazzi of the University of Milan and his colleagues describe an upturn in the incidence of particular cancers among the Seveso population. People living in the second most contaminated area, called zone B, were nearly three times more likely to acquire liver cancer than was the general population. In this same cluster, a form of myeloma occurred 5.3 times more often among women; among men, some cancers of the blood were 5.7 times more likely.

The researchers did not find a greater number of the cancers in the most polluted area, a fact Bertazzi anticipated. The small group of people most affected moved immediately, so their exposure was short, Bertazzi says. Those in zone B had lower, prolonged exposure.

These findings are not the first to associate dioxin with cancer in humans; over the years, various studies have found evidence for and against such a link. The Seveso study is significant because this population has been well monitored and because new techniques have made blood levels of dioxin easy to measure—a crucial factor in accurately determining exposure. Although Bertazzi has based his findings on extrapolations from soil data, the investigator says analyses of the blood samples correspond to his estimates.

The Seveso study may be important even for what is absent from it. Bertazzi notes that the occurrences of breast cancer and endometrial cancer are below normal. "What is remarkable about these findings is that they reflect animal data almost perfectly," comments Ellen K. Silbergeld, a toxicologist at the

University of Maryland and a staff scientist at the Environmental Defense Fund. Both cancers are thought to be induced by estrogen. Because dioxin functions in part as an antiestrogen, it may work to protect against such cancers, Silbergeld explains.

The Seveso findings also come at a time when information about the molecular effects of dioxin have begun to accumulate. Scientists understand that dioxin—in particular, 2,3,7,8-tetrachlorodibenzo-*para*-dioxin, the most potent of the 75 types of dioxin—binds to an intracellular receptor. The dioxin-laden receptor then joins with a transporter that shuttles the complex to a cell's nucleus and activates an enzyme, cytochrome p450. "When the complex interacts with the DNA, it disrupts the chromosome structure," says James P. Whitlock, Jr., a pharmacologist at Stanford University. The resulting changes in gene expression have led investigators to postulate that dioxin promotes cancer caused by another substance.

Other studies have suggested that dioxin functions as a hormone and affects the immune system and the reproductive tract. Sherry E. Rier of the University of South Florida reported in *Fundamental and Applied Toxicology* that dioxin is associated with endometriosis in rhesus monkeys. The National Institute of Environmental Health Sciences (NIEHS) is studying the same association in women. Most U.S. occupational studies of dioxin—which constitute the bulk of such research—have not examined its impact on women, who seldom encounter the compound in the workplace.

Richard E. Peterson of the University of Wisconsin and others have also found that dioxin can cause neurobehavioral changes in rats and can alter reproductive tract development. Similar findings have been seen in a population poisoned by a dioxin analogue in Taiwan. Boys who were exposed in utero have smaller penises than do unexposed boys.

"Dioxin is a very potent growth regulator," notes Linda Birnbaum, a toxicologist at the Environmental Protection Agency. "It has many different effects on many different organ systems—at different stages of development." The EPA is evaluating the new data as it continues its reassessment of dioxin. The agency is expected to issue its review this year.

And Bertazzi's paper is not the last word from Seveso. George W. Lucier, a biochemist at the NIEHS, and others are looking at the induction of cytochrome p450 in the Seveso residents to see if it is associated with the development of cancer.

—*Marguerite Holloway*



PROFILE: ALBERT EINSTEIN

Keyhole View of a Genius

Albert Einstein scholars have long been aware of troubled and troubling aspects of the great physicist's life. His first marriage, strongly disapproved of by his family, ended in divorce. The child of this union was put up for adoption. Letters and other documents in *The Collected Papers of Albert Einstein*, a compendium of Einstein's papers, published by Princeton University Press, contain hints of infidelity. Yet his scholars and biographers have focused on his work or turned discreetly away from this aspect of his life.

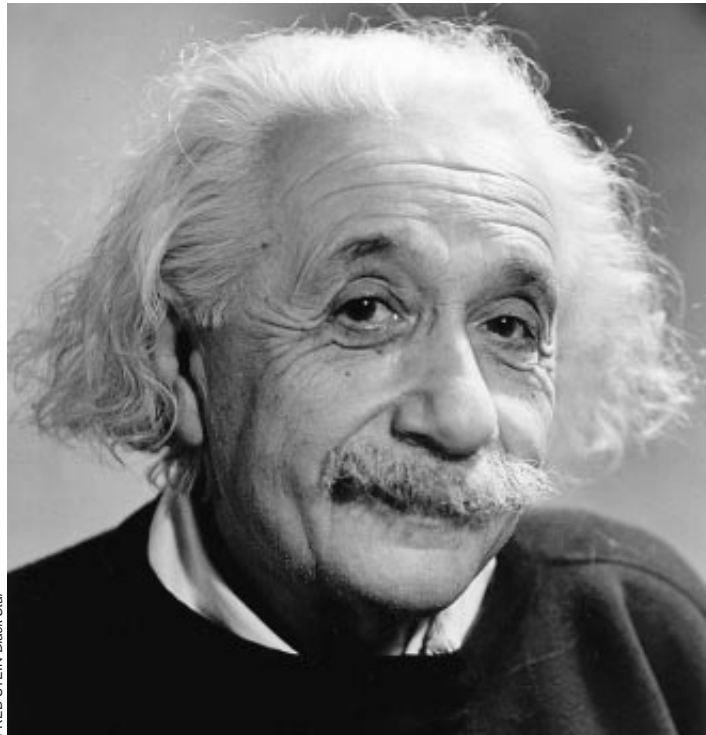
In doing so, they have left the field open. And Fleet Street abhors a vacuum. So instead of the kind of scholarship that would provide us with a rounded picture of this complicated, powerfully gifted human being, we have *The Private Lives of Albert Einstein*. In the book, which was published last August in Britain by Faber & Faber, two English journalists, Peter Highfield and Paul Carter, report the results of a quick foray they have made into *The Collected Papers*. (To date, three volumes have appeared; two more are expected.)

Highfield and Carter's booty consists of a series of letters, which they have fleshed out with interviews and—where evidence fails—with their own speculation. Using such materials, the authors have created a portrait of a man of physical passion who conducted a complicated romantic life as he revolutionized the foundations of contemporary physics and cosmology. St. Martin's Press will publish the book in the U.S. this spring.

By the time Einstein left war-torn Europe to take his place as a cultural icon in the U.S., he had already finished the work that established him as a seminal figure in modern physics. For scientists, the work counts above all; Einstein, the man, comes second. Einstein would have approved of these priorities. His

highest praise, once given in a generous moment to his eldest son, was to possess "the ability to rise above mere existence by sacrificing one's self through the years for an impersonal goal."

Einstein's own mere existence, as seen by Highfield and Carter, consists of a collage of personas only faintly recognizable to readers of previous biographies. First we learn that Einstein was an alienated and overmothered youth.



FRED STEIN Black Star

Then we meet the adolescent Einstein, bursting with libido: "a handsome teenager exuding casual charisma" who possessed "masculine good looks," a "raffish" mustache and a "muscular and quite powerful" physique. (The genial gnome of the classic portrait is also a myth: even in old age, Einstein was a physically robust man.)

As a youth, Highfield and Carter say, Einstein was both passionate and calculating in his handling of women. He pens a love poem to one teenage acquaintance—"...a kiss on your tiny little mouth..."—while reassuring wife-to-be Mileva of his continuing devotion.

According to Highfield and Carter, *The Collected Papers* reveals a dark,

perhaps violent, side of Einstein that appears several years into his first marriage, particularly after his 1905 papers on special relativity begin to attract recognition. Einstein and Mileva argue fiercely over his contact with other women, and Einstein, in letters to his friend Michele Besso, attributes her jealousy to a pathological flaw typical of a woman of such "uncommon ugliness." One day Lisbeth Hurwitz notes in her diary that she has seen Mileva's face badly swollen. The authors leave the reader to decide whether the cause

was a blow or a toothache. Highfield and Carter note that before the breakup of Einstein's first marriage, the young physicist lived with Elsa Einstein, a cousin, in Berlin, leaving his wife and their two children in Zurich, unable to pay the rent.

As the Highfield and Carter narrative unfolds, Einstein's misogyny increases as does his fame. He was, for example, a friend of the renowned Franco-Polish scientist Marie Curie. He nonetheless refers in a letter to Elsa to Curie's "severe outward aspect" and says she has "the soul of a herding." He also spins theories to explain what he regards as the inherent inability of women to think great scientific thoughts.

Eventually he divorces Mileva and marries Elsa,

but, the authors claim, the philandering goes on. At least one woman, a young blonde, visits him regularly at his summer house in Berlin, where they take boating excursions while Elsa consoles herself with pastries and cakes, according to a maid whom the authors interviewed. In another anecdote, as related by the physicist's friend Janos Plesch, Einstein stops one day to ogle a woman kneading bread.

Every now and then the amorous Einstein portrayed by Highfield and Carter does pause to do a bit of physics. He also shows a few glimpses of compassion to his loved ones. But on the whole, he is Mr. Hyde to the Dr. Jekyll of popular Einstein myth. "We wanted

to provide an antidote to the previous biographies," Highfield explains.

What can be gained by examining this "mere existence" of Einstein's? "You can't get a feeling for what Einstein was like by reeling off his scientific achievements," Highfield observes. Some scientists who knew Einstein disagree. "I was somewhat unhappy at the publication of all this material," says Peter Bergmann, Einstein's collaborator during his days at the Institute for Advanced Studies in Princeton, N.J. "Being dead, you don't give up your claims to privacy," Bergmann declares.

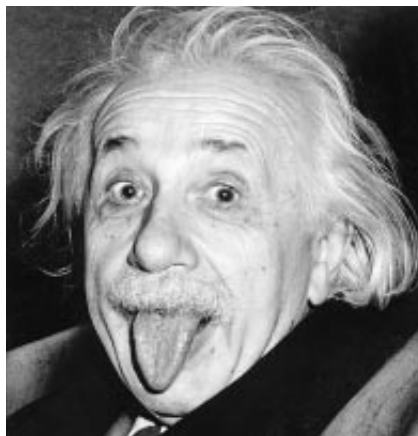
Bergmann places himself firmly in the camp of Einstein's executors: his former secretary Helen Dukas and friend Otto Nathan. Einstein, having kept his two sons from his first marriage, Hans Albert and Eduard, at an emotional distance, enlisted the possessive Dukas as "mother protector" after the death of Elsa. In his will, he left her and Nathan in charge of his literary legacy. They guarded it vigorously, preventing in 1958 the publication of a manuscript written by Frieda Einstein, Einstein's daughter-in-law, that was based in part on letters from Mileva. Did such policing keep valuable truths from scholars? "Historians may think so," Bergmann asserts, "but I have my doubts."

Nevertheless, after the deaths of Dukas and Nathan, the letters found their way to the Hebrew University of Jerusalem, and *The Collected Papers* project was begun. Bergmann was on the losing side of heated arguments among fellow scientists advising the publishers on whether to include particularly intimate letters. But John Stachel, director of the Center for Einstein Studies at Boston University and former editor of *The Collected Papers*, calls the book well documented and serious, even though he disagrees with many of Highfield and Carter's conclusions. "If you think Einstein was a plaster saint," he says, "you'll be upset."

Abraham Pais, author of *Subtle is the Lord*, an Einstein biography concerned mainly with Einstein's scientific achievements and regarded by many physicists as definitive, agrees with Stachel about the need to publish the archives in their entirety. But the relentless focus on Einstein's romantic and erotic behavior in *Private Lives* makes him seethe. "It could be worse," he says, "but not much. So [Einstein] had a few extramarital affairs. That happens in the best of families. The book's emphasis is wrong." Sir Martin Rees, a professor of astronomy at the University of Cambridge, entertains similar sentiments: "It's entirely appropriate to learn everything you can about somebody you're

writing about. But at all points, [Highfield and Carter] place the worst possible construction on Einstein's motives."

Highfield and Carter indeed go to considerable lengths to paint Einstein in the worst light possible. *Private Lives* relies almost exclusively on circumstantial evidence and indirect references to support many of its conclusions. This practice holds especially true for many of the claims about Einstein's philan-



dering. Take the case of Grete Markstein, a Berlin actress who claimed in 1935 to be Einstein's long-lost daughter. The archives contain plain evidence that Einstein sired a daughter by Mileva before their wedding, whom the couple are believed to have put up for adoption. Although Einstein dismissed Markstein's claim out of hand, he took the trouble to have his secretary hire a detective to check out her story. It turned out to be untrue, but documentary evidence suggests that three years earlier Einstein made a payment of 80 marks to Markstein for "semiofficial" services. Again, the authors point the reader toward an unseemly conclusion.

Even Einstein's efforts to intervene in the lives of his children sound like tales from a stag party. According to the authors, Einstein wrote to Mileva about his disapproval of their son Hans Albert's bride-to-be. Einstein suggests that his son's choice of a domineering woman is the result of sexual inhibitions. Allegedly, Einstein proposes that the son be sent to a pretty 40-year-old woman of the physicist's acquaintance for unspecified remedial instruction. And, shades of Woody Allen, the authors point out that in Einstein's later years his stepdaughter, Margot, appeared with him almost everywhere he went—far more than did his wife, Elsa.

Although he admits that the archives provide no hard evidence for many of his and Carter's contentions, Highfield stands by them. Because people were not in the habit in the early part of the

century of recording intimate items in their letters, Highfield believes, he and Carter had to rely on indirect references. "You have to look at the overall accumulation of these details," he says.

Jürgen Renn, a physicist who until recently participated in the preparation of *The Collected Papers* and who is now director of the Max Plank Institute for the History of Science in Berlin, argues that the personal details in *Private Lives* actually offer some insight into the creative process behind Einstein's achievements. "You can't understand the peculiar combination of what he did and when he did it without knowing about his personal life," he says.

That Einstein and Mileva lived like "bohemian outsiders" in the period before 1905 had an impact on Einstein's theory of special relativity, Renn contends. His marriage to Mileva estranged Einstein from his family, and he was having trouble finding a job. The physicist's arrogance and rebelliousness, coupled with his relationship with Mileva, "gave him the courage to take up [scientific] issues that he wouldn't have taken up otherwise," Renn says. That assumption goes far toward explaining why Einstein, in his correspondence with Mileva, referred in 1901 to "our work on relative motion." By the same token, Einstein's later move to Berlin constituted something of a return to the "inside"—to his new job at the center of the physics establishment and to the good graces of his family. Mileva no longer suited his changed sensibility.

So far few physicists seem to have actually read *Private Lives*. Roger Penrose, Rouse Ball Professor of Mathematics at the University of Oxford and author of *The Emperor's New Mind*, does not put the book high on his reading list, although he is keen to peruse the letters in *The Collected Papers* volumes. David Robinson, a professor of mathematics at King's College, London, comments that "most working physicists like me will wait for the paperback version." George P. Efstathiou, Savilian Professor of Astronomy at Oxford, says the book has given him insight into Einstein's character: "It's not the sexual misdemeanors that interest me but rather Einstein's independence from authority in his younger years."

Efstathiou may be on the right track. Once the salacious curiosity has been satisfied, Pais, Bergmann or other serious scholars face the fascinating challenge of exploring the complex personality that *The Collected Papers* reveals. And sociologists or other soft scientists may want to examine society's need for idols, a need history seems ever ready to frustrate. —Fred Guterl, London

Wetlands

These havens of biodiversity are often endangered because they can be hard to identify. Understanding their variable characteristics can lead to more successful conservation efforts

by Jon A. Kusler, William J. Mitsch and Joseph S. Larson

Variably dry, wet or anywhere between, wetlands are by their nature protean. Such constant change makes wetlands ecologically rich; they are often as diverse as rain forests. These shallow water-fed systems are central to the life cycle of many plants and animals, some of them endangered. They provide a habitat as well as spawning grounds for an extraordinary variety of creatures and nesting areas for migratory birds. Some wetlands even perform a global function. The northern peat lands of Canada, Alaska and Eurasia, in particular, may help moderate climatic change by serving as a sink for the greenhouse gas carbon dioxide.

Wetlands also have commercial and utilitarian functions. They are sources of lucrative harvests of wild rice, fur-bearing animals, fish and shellfish. Wetlands limit the damaging effects of

waves, convey and store floodwaters, trap sediment and reduce pollution—the last attribute has earned them the sobriquet “nature’s kidneys.”

Despite their value, wetlands are rapidly disappearing. In the U.S., more than half of these regions in every state except Alaska and Hawaii have been destroyed. Between the 1950s and the 1970s more than nine million acres—an area equivalent to the combined size of Massachusetts, Connecticut and Rhode Island—were wiped out. Some states have almost entirely lost their wetlands: California and Ohio, for example, retain only 10 percent of their original expanse. Destruction continues today, albeit at a slightly reduced rate, in part, because there are fewer wetlands to eliminate. No such numbers are available internationally, but we estimate that 6 percent of all land is currently wetlands.

The extensive losses can generally be attributed to the same feature that makes wetlands so valuable: their ever changing nature. The complex dynamics of wetlands complicate efforts to create policies for preserving them. Their management and protection must incorporate a realistic definition, one that encompasses all these intricate

JON A. KUSLER, WILLIAM J. MITSCH and JOSEPH S. LARSON work on aspects of wetland management and ecology. Kusler, who has advised many state and federal agencies on water resource policy, is executive director of the Association of Wetland Managers. Professor of natural resources and environmental science at Ohio State University, Mitsch has conducted extensive research on wetlands restoration and ecosystem modeling. Larson is professor at and director of the Environmental Institute at the University of Massachusetts at Amherst. He has studied, among other topics, the behavior of beavers and the assessment of freshwater wetlands.

FLOODING IN THE MIDWEST left thousands of houses submerged—including these along the Missouri River—and powerfully demonstrated the dangers of destroying wetlands. When undisturbed, wetlands can absorb excess floodwater. Development, however, can reduce or eliminate this capability.



ecosystems—from marshes, bogs and swamps to vernal pools, playa lakes and prairie potholes. If scientists can better clarify and communicate to the public and to policymakers the special characteristics of wetlands as well as their economic and ecological importance, perhaps those that do remain will not disappear.

Over the years, researchers and government agencies have developed many definitions of wetlands. All share the recognition that wetlands are shallow-water systems, or areas where water is at or near the surface for some time. Most descriptions also note the presence of plants adapted to flooding, called hydrophytes, and hydric soils, which, when flooded, develop colors and odors that distinguish them from upland soils.

Wetlands can be found in diverse to-

pographical settings. They arise in flat, tidally inundated but protected areas, such as salt marshes and mangrove swamps. Wetlands exist next to freshwater rivers, streams and lakes and their floodplains (such areas are often called riparian). In addition, they form in surface depressions almost anywhere. Such wetlands comprise freshwater marshes, potholes, meadows, playas and vernal pools where vegetation is not woody, as well as swamps where it is. Wetlands can also flourish on slopes and at the base of slopes, supplied by springs, and as bogs and fens fed by precipitation and groundwater. Finally, they can occur in cold climates where permafrost retains water and low evaporation rates prevail.

Although the kinds and locations of wetlands vary greatly, fluctuating water levels are central to all of them. Water rises or falls in accordance with tides,

precipitation or runoff; the activities of humans and other animals can also determine water levels. The extent of the fluctuation is often very different from site to site. In the salt marshes of the northeastern U.S. and eastern Canada, daily tides may bring about shifts of 10 feet or more in water level. Other regions undergo even more extreme changes. For example, rainfall can cause the Amazon River to rise 25 feet during a season and invade neighboring wetlands [see “Flooded Forests of the Amazon,” by Michael Goulding; *SCIENTIFIC AMERICAN*, March 1993]. In the prairie potholes of the Midwest, groundwater or melting snow may alter water levels by four or five feet over several years.

Even when levels fluctuate dramatically, these systems can adjust so that they sustain little permanent damage. Indeed, the very existence of some wetlands is related to the ravages of hurri-



canes, floods and droughts. Most wetlands along rivers and coastlines as well as those that formed in depressions in the landscape are long-lived precisely because of events that people consider economically devastating. Raging fires burn excess deposited organic matter

and recycle nutrients. Hurricanes and high-velocity floods scour sediments and organic matter, removing them from wetlands or creating wetlands nearby. Droughts temporarily destroy hydrophytic vegetation and allow oxidation and compaction of organic soils.

This anomalous feature of wetlands—the way that short-term destruction ensures long-term gain—is poorly understood by the general public. Much of the press coverage of Hurricane Andrew and its impact on the Florida Everglades illustrates this fact. Although

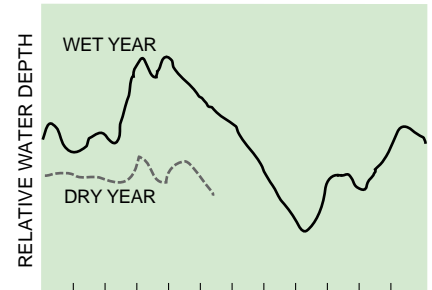
The Fluctuating Water Levels of Wetlands

Wetlands are often as different in their appearance and in the species they host as they are in the range of saturation they experience in the course of a year or a season.

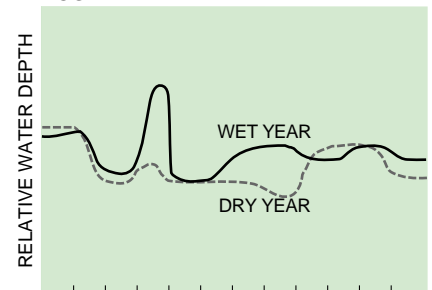
Their topographical variety and the complexity of their hydrology have made some wetlands difficult to identify and, hence, difficult to preserve.



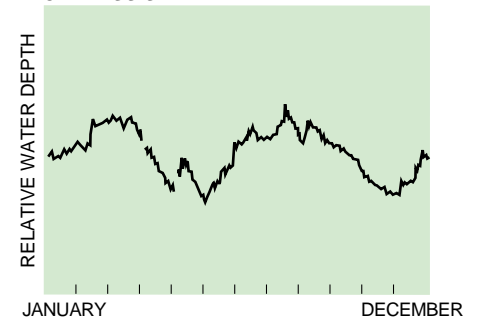
PRAIRIE POTHOLE



BOG



CYPRESS SWAMP



the damage was serious, the ecosystem and others like it have survived thousands of such cataclysms. Some researchers have suggested that trees in the coastal mangrove swamps reach maturity at about 30 years of age, a periodicity that coincides almost perfectly

with the frequency of hurricanes in the tropics.

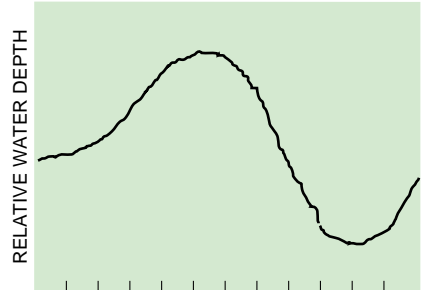
Misunderstanding has also led to many well-intentioned proposals to stabilize water levels in wetlands. The flooding along the Mississippi, Missouri and other rivers last summer was especially

severe because wetlands had been destroyed as people built on them. These ecosystems could no longer serve to absorb floodwaters.

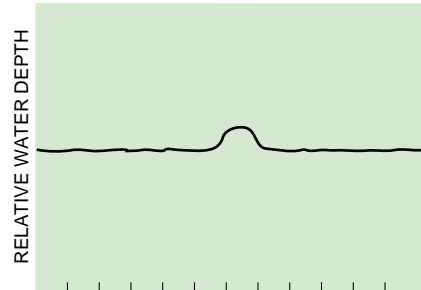
Of course, the levels of many bodies of water rise and fall. Lakes and streams are occupied by plants and animals



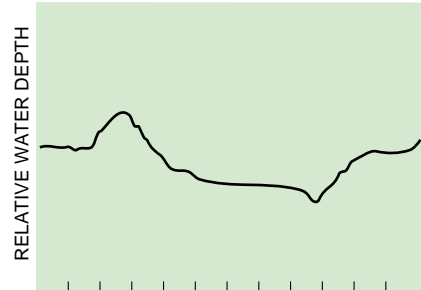
TROPICAL FLOODPLAIN



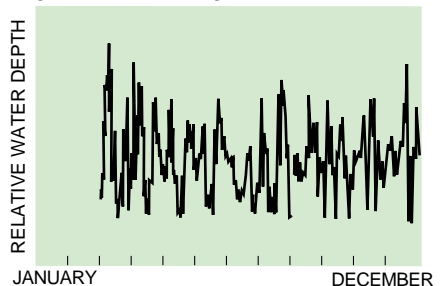
TUNDRA

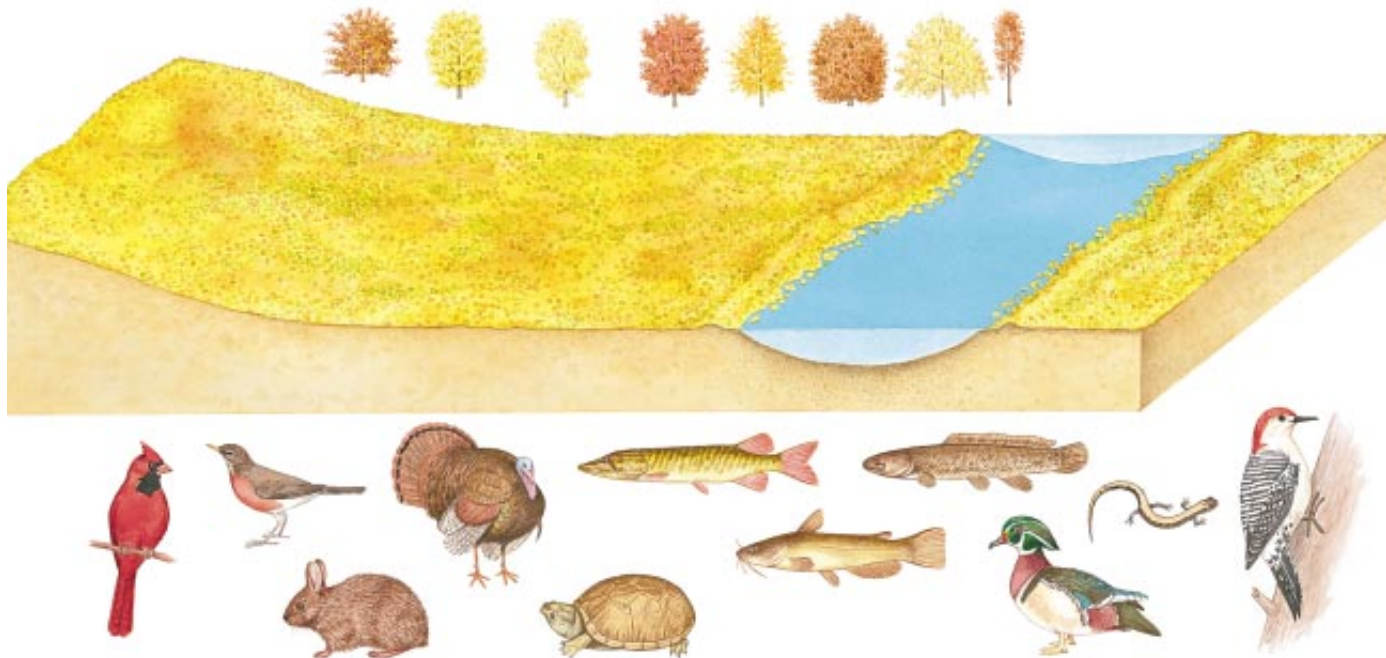


FRESHWATER MARSH



SALTWATER MARSH





BOTTOMLAND HARDWOOD WETLANDS that occur in the major river basins of the southeastern U.S. have two very distinct hydroperiods, or periods of inundation. During the dry season (*left*), fish species such as the yellow bullhead stay in the channel, whereas animals and birds move through out all zones of the region. But during the flooded period

that are adapted to a permanently watery environment—even temporary dry spells could kill them. In contrast, a wetland encompasses an array of shallow-water and saturated soil environments that possess some elements of a terrestrial system and some of an aquatic system. Because water levels rise and fall continuously, portions of wetlands—and, in some cases, entire wetlands—at times resemble true aquatic systems, at times terrestrial systems and at times intermediate systems. Plants, animals and microbes are constantly adapting and changing.

Wetlands also differ from deep-water aquatic systems in their sensitivity to the effects of water-level changes. A one-foot change in the level of a lake or a river brings about little difference in a system's boundaries or functions. But an equivalent change in a wetland can significantly affect both. Certain wetland vegetation—sedges, grasses or floating plants—often grows in one location during a wet year, another location during an intermediate year and not at all during a dry year. Thus, cycles of plant growth can change over time. As a result, the kinds of animals that frequent a wetland will also vary.

Such shifts explain the immense biodiversity of wetlands. Alterations in their water levels give rise to a series of ecological niches that can support terrestrial, partially aquatic and fully aquatic plants and animals. In addition, vertical gradients caused by differing

depths of water and saturation create further environmental variation. Wetlands essentially borrow species from both aquatic and terrestrial realms.

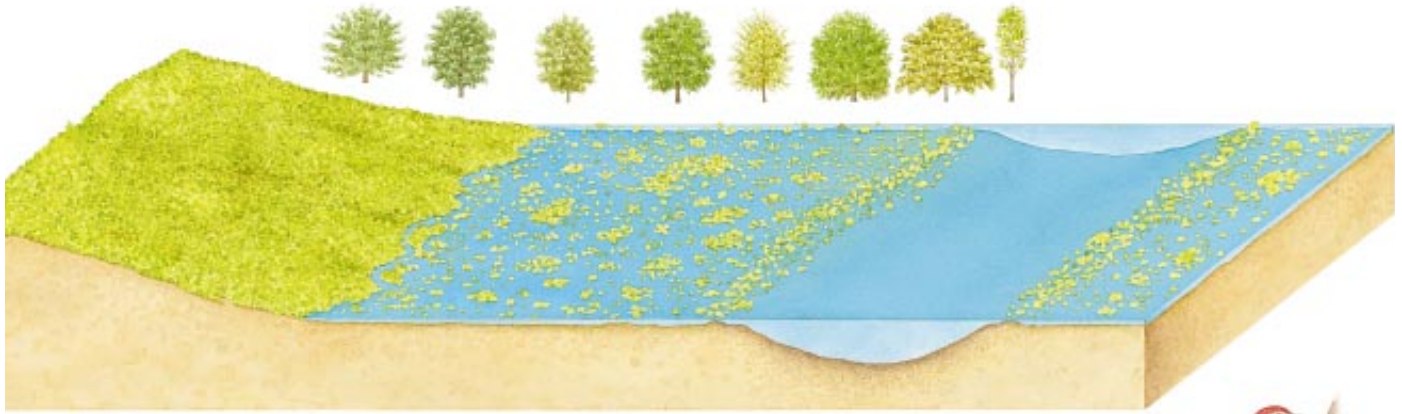
Even a temporary niche can be crucial to the nesting, spawning, breeding or feeding patterns of a particular species. Short-legged birds such as green-backed herons and limpkins feed along shallow-water shorelines. Longer-legged species, including egrets and great blue herons, feed in deeper water. Swimming waterfowl such as mallards, coots and purple gallinules feed in the deepest open water. Shifts in water levels serve to trigger nesting by wood storks in Florida and breeding by ducks in prairie potholes.

Rising and falling water levels not only influence the internal character of a wetland, but they also link wetlands to one another and to other aquatic systems. Because of their sensitivity to water levels, wetlands are highly dependent on the quantity and quality of water in their immediate area. This fact is particularly true for isolated or small wetlands. In such terrain, rain, local runoff and the aquifer are the only sources of water. Wetlands bordering major lakes and streams may be less sensitive to such natural changes. They rely on the levels in adjacent water bodies that, in turn, depend on precipitation in larger watersheds. Coastal wetlands are also somewhat more resilient since levels depend on the tides.

Such associations with the neighboring environment are critical to wetland functions. Wetlands can serve as reproductive or feeding sites for some species only if they are connected with other waterways. Moreover, the incoming water brings nutrients and sediments that can make the system more productive. The wetlands then cleanse these waters by retaining sediments as well as phosphorus and other chemicals. Pollutants such as nitrogen can be turned into harmless gases by the aerobic and anaerobic bacteria found there.

Clearly, the dependence of many wetlands on contiguous water systems makes them especially vulnerable to even minor human activity. Development in watershed areas and the pumping of groundwater can disrupt or destroy them. Landfills, dikes or other measures that isolate wetlands from nearby wetlands or waters can reduce their ability to provide flood storage, water purification and habitats.

Barriers also can prevent wetland plants and animals with highly sensitive aquatic tolerances from migrating up and down gentle slopes. Without sufficient room to move, wetlands themselves may temporarily or permanently disappear. Some—including headwater riparian wetlands, depressional wetlands and slope wetlands—are particularly prone to such interference. A seawall or a dike at the landward boundary of a salt marsh can prevent the inland migration of the marsh when the sea



(right), the crucial role of the wetland as spawning ground and nursery becomes evident. The fish move into the inundated forest, where they spawn and feed; wood ducks fly

into the area to nest. Many other creatures move upland to dry ground. The bottomland hardwood plants and animals are thus adapted to both the dry and the wet periods.

level rises. Indeed, such diking currently threatens, rather than helps, many coastal areas.

Increased amounts of sediment, nutrients and pesticides from watersheds undergoing development can drastically alter the biological makeup of a wetland and overload its ability to purge pollutants if they are added beyond the wetland's ability to assimilate them. Such additions can even destroy a wetland in a short time. Isolated wetlands arising in topological depressions are quite vulnerable because they are not periodically purged of sediment by storms or high-velocity river flows.

Many pothole and kettle-hole wetlands in the northern American states and the southern parts of Canadian provinces are at just such risk. Most wetlands in these regions were created 8,000 to 12,000 years ago by the retreat of the glaciers. As blocks of ice in glacial outwash and till (the assemblage of rocks, boulders and clay that rides along with the glacier) melted, pothole depressions were formed. The deeper ones became lakes; the shallow ones, wetlands. In presettlement times, heavily vegetated surroundings contributed small amounts of sediment and nutrients to these wetlands. But the clearing of land increased this influx of sediment, which continues to build up because the ecosystems lack effective flushing mechanisms.

Ironically, decreased sediment from dams and reservoirs along rivers and

streams threatens other wetlands. In the Mississippi Delta, levees have prevented loads of sediment from being deposited—to the point that marshes can no longer build up at a rate equal to sea-level rise and land subsidence. The result is a massive loss, an estimated 25,000 acres of marsh every year. Watershed development and diversions that decrease the freshwater flow of rivers similarly threaten many estuarine wetlands by reducing the quantity of freshwater and increasing salinity.

It is not difficult to see how fluctuating water levels and the intricate relations between wetlands and human development pose serious challenges to any simple wetland policy. Highly generalized rules are often insensitive to the physical characteristics and dynamics of wetlands.

To some extent, the battle over wetlands has been a conflict between conservation and development. There is hardly a farmer, developer or shopping-mall builder in the U.S. who is not familiar with wetlands. The debate has pivoted around the problem of devising management strategies that provide certainty for developers while protecting the ecological features of wetlands. Fluctuating water levels and the sensitivity of wetlands to these changes as well as the dependence of wetlands on the surrounding landscape must consistently be taken into account.

Landowners understandably want to

know the exact effect of wetland regulations when they construct a house or road. They want to know what activities will be allowed in which areas under what conditions. They want to be able to compensate for wetland losses at one site by restoring wetlands at other locations. And they want hard and fast rules, without surprises.

This need has led to proposals to take wetland policy out of the hands of the scientists and to establish simplistic rules through legislative fiat. Such attempts include congressional bill HR 1330, co-sponsored by 170 members of the House in 1992 and 100 members in 1993, which provides an example of science and legislation in conflict. The bill would require that hydric vegetation be present in every wetland. It also stipulates that wetlands be classified according to a once-and-for-all determination of a wetland's value or function.

In essence, HR 1330 treats wetlands like static water systems. (A similar problem of failing to recognize wetlands as a dynamic system was seen in the fall of 1991, when the U.S. administration tried and failed to redefine wetlands.) Moreover, the proposal would allow a landowner to select the time of year during which to decide whether or not a particular area constitutes a wetland. Because such hydric plants are missing at one time or another from most wetland sites, provisions of this kind could be used to define most wetlands out of existence.



FLORIDA EVERGLADES appeared to be severely damaged by Hurricane Andrew, which ripped through the region in 1992. Yet contrary to public perception, the wetlands that make up

the Everglades rely on such storms for their survival. Gale-force winds remove excess organic matter and sediment that are suffocating the ecosystem.

The bill would require that federal agencies document 21 days of inundation or saturation for all wetlands. This artificial standard would be impossible to meet because water-level records are rarely available, and fluctuations are extremely difficult to predict. The expense of using modeling to foresee water levels is prohibitive: one study to determine the probability of a 100-year, or extremely rare, flood on about half the nation's floodplains cost more than \$870 million.

Finally, the bill, which would allow for compensating the loss of one wetland by preserving another—called mitigation banking—ignores the tight associations between certain wetland functions and their watershed. A wetland's ability to control floodwater or maintain water quality can be seen immediately downstream. But, under the bill, downstream landowners are not compensated for the fact that their wetlands can no longer fulfill these functions. Further, because of their surroundings, two wetlands of similar size in different locations may have distinctly different attributes, functions and therefore value.

Scientifically sound management of wetlands that satisfies everyone is not easy to achieve, but there are signs of hope. In the past decade, investigators have learned much

about defining and managing wetlands as dynamic features in the landscape. This knowledge could form the basis of a workable policy.

Recognizing the role of fluctuating water levels and the interrelation of the landscape is a first step. Water levels vary within relatively well defined ranges in most wetlands and can therefore provide a foundation for definition and regulation. Soil and geologic information can be gathered to identify long-term shifts. Other criteria can help indicate altered or managed wetlands as well as those that are infrequently flooded. It is also important to consider the immediate landscape when the wetland is being evaluated.

In the future, natural processes should be preserved as much as possible. In general, people have attempted to control the rise and fall of rivers by building dams. When such fluctuations cannot be maintained, remedial management should be undertaken to simulate natural hydrologic pulses.

Regional watershed analyses that address not only present but future situations can help delineate wetlands. These analyses can form the foundation for planning and regulation. At the same time, protection of these systems can be integrated into broader land-use policies—including the management of water supplies and of floodplains, storm water and pollution.

Such scientifically sound policies have been implemented in many countries. In 1971 the Ramsar Convention called for the protection of wetlands and for the formulation of national plans to use them wisely. Today 37 million hectares at 582 sites have been designated as Ramsar sites—including 1.1 million hectares in the U.S. Nevertheless, only 74 nations have joined the convention.

Because of their special characteristics, wetlands pose difficult but not insurmountable challenges in terms of protection and restoration. If we recognize these features and incorporate them into policies at all levels of government, we can save the remaining wetlands, from the tropics to the tundra.

FURTHER READING

WETLAND CREATION AND RESTORATION: THE STATUS OF THE SCIENCE. Edited by Jon A. Kusler and Mary E. Kentula. Island Press, 1990.

WETLANDS: A THREATENED LANDSCAPE. Edited by Michael Williams. Basil Blackwell, 1991.

WETLANDS. Edited by M. Finlayson and M. Moser. Facts on File, 1991.

WETLANDS. William J. Mitsch and James G. Gosselink. Van Nostrand Reinhold, 1993.

WETLANDS IN DANGER: A WORLD CONSERVATION ATLAS. Edited by Patrick Dugan. Oxford University Press, 1993.

The Search for Strange Matter

Between nucleus and neutron star stretches a desert devoid of nuclear matter. Could strange quark matter fill the gap?

by Henry J. Crawford and Carsten H. Greiner

For some years, physicists have enjoyed toying with a particularly intriguing puzzle. Protons and neutrons readily form either tiny clumps of matter (the various atomic nuclei) or very large clumps of matter (neutron stars). Yet between the invisible nucleus and the ultradense neutron star (really a vast nucleus that is some 11 kilometers or more in circumference), no form of nuclear matter has been detected. What is going on here? Do the laws of physics as we know them forbid nuclear particles from assembling themselves into objects that could fill this “middle” range? Or is this nuclear desert actually filled with new forms of matter, different in structure from ordinary nuclear matter, that investigators have failed to find?

In fact, the theory that embodies our current understanding of physics, the Standard Model, seems to be consistent with the existence of new forms of nuclear matter that might populate the desert. And if the Standard Model is right, the detection of such matter could solve a major cosmological mystery: the nature of the “missing” matter, thought to account for 90 percent of the observable universe. This is a

prize worth winning. So, in an experiment at Brookhaven National Laboratory, we, along with many collaborators from other research institutions, are searching for evidence of the existence of this form of nuclear matter that might fill the void.

According to the Standard Model, all matter consists of quarks. Six varieties of these particles exist, grouped into three sets of twins: “up” and “down,” “strange” and “charm,” “top” and “bottom” (or “truth” and “beauty”). All but one (the top quark) have been observed. Only two kinds of quarks figure in our daily lives: up and down.

A proton consists of two up quarks (each of which has a fractional charge of $+2/3$) and a down quark (whose charge is $-1/3$). Two down quarks ($-1/3$, $-1/3$) and an up quark ($+2/3$) make up the neutron. The other varieties, or flavors, have thus far been found only within short-lived particles. Recent theoretical calculations raise the possibility that the two flavors of quark found in ordinary matter combined with a third flavor, the strange quark, could form stable entities. Such strange quark matter could easily assemble itself into entities whose sizes fall between that of the nucleus and the neutron star.

To understand how strange quark matter might materialize, we must go

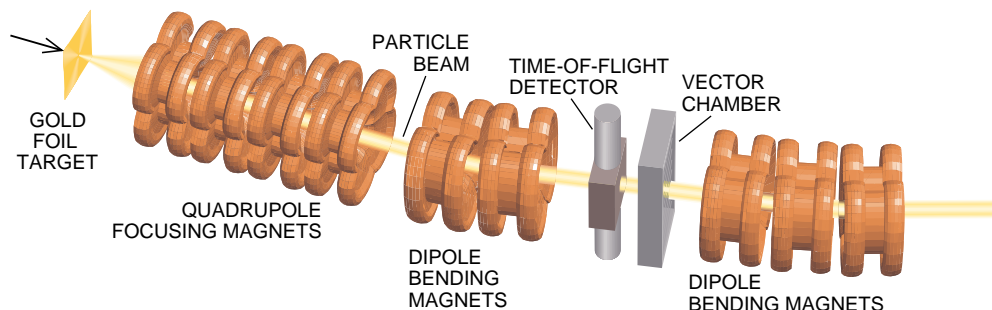
deeper into the Standard Model. Protons, neutrons and other particles formed from quarks are called hadrons (from the Greek *hadros*, meaning robust). For simplicity, physicists often model hadrons as tiny “bags” in which the quarks freely roam but from which they cannot escape. All known hadronic particles consist of either bags harboring three quarks—the baryons—or a quark and an antiquark—the mesons. (Each quark, like every elementary particle, has an antimatter twin.)

Quarks inside the bag can change their identity via the actions of the weak force, which is responsible for the beta decay of nuclei. The weak force changes the down quarks into up quarks. A neutron (up quark, down quark, down quark, or udd) can become a proton (up quark, up quark, down quark, or uud) when the weak force changes one of its down quarks to an up (an electron and an antineutrino are also emitted in the process). The weak force can also change the strange quark into a down quark. This effect explains why particles containing strange quarks, such as the lambda (a baryon containing an up, a down and a strange quark) or the *K*'s (mesons containing an anti-strange quark paired with either an up or a down quark), are not stable.

We are closing in on the prize. We do

HENRY J. CRAWFORD and CARSTEN H. GREINER are collaborators in an investigation at Brookhaven National Laboratory that aims to produce and detect strange quark matter. Crawford is a research scientist at the Space Sciences Laboratory at the University of California, Berkeley, where he received his doctorate in 1979. He has used satellites and particle accelerators to pursue his primary research interest in nuclear astrophysics. Greiner received his Ph.D. from the University of Erlangen-Nürnberg in Germany in 1992. He is currently an Alexander von Humboldt Fellow and a visiting assistant professor at Duke University, continuing his research on theoretical aspects of nuclear matter under extreme and nonequilibrium conditions.

DETECTOR at Brookhaven National Laboratory is part of an experiment to create and find strange matter. The pink cylinders are Cerenkov counters, which detect fast-moving charged particles. Similar experiments are currently under construction at other laboratories around the world.

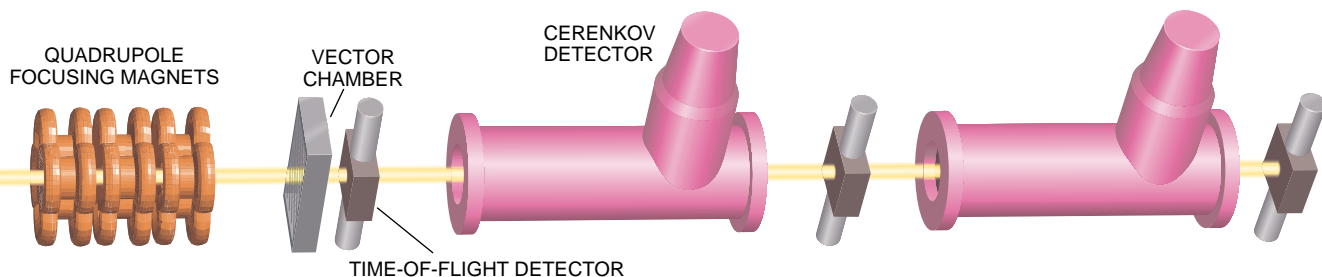
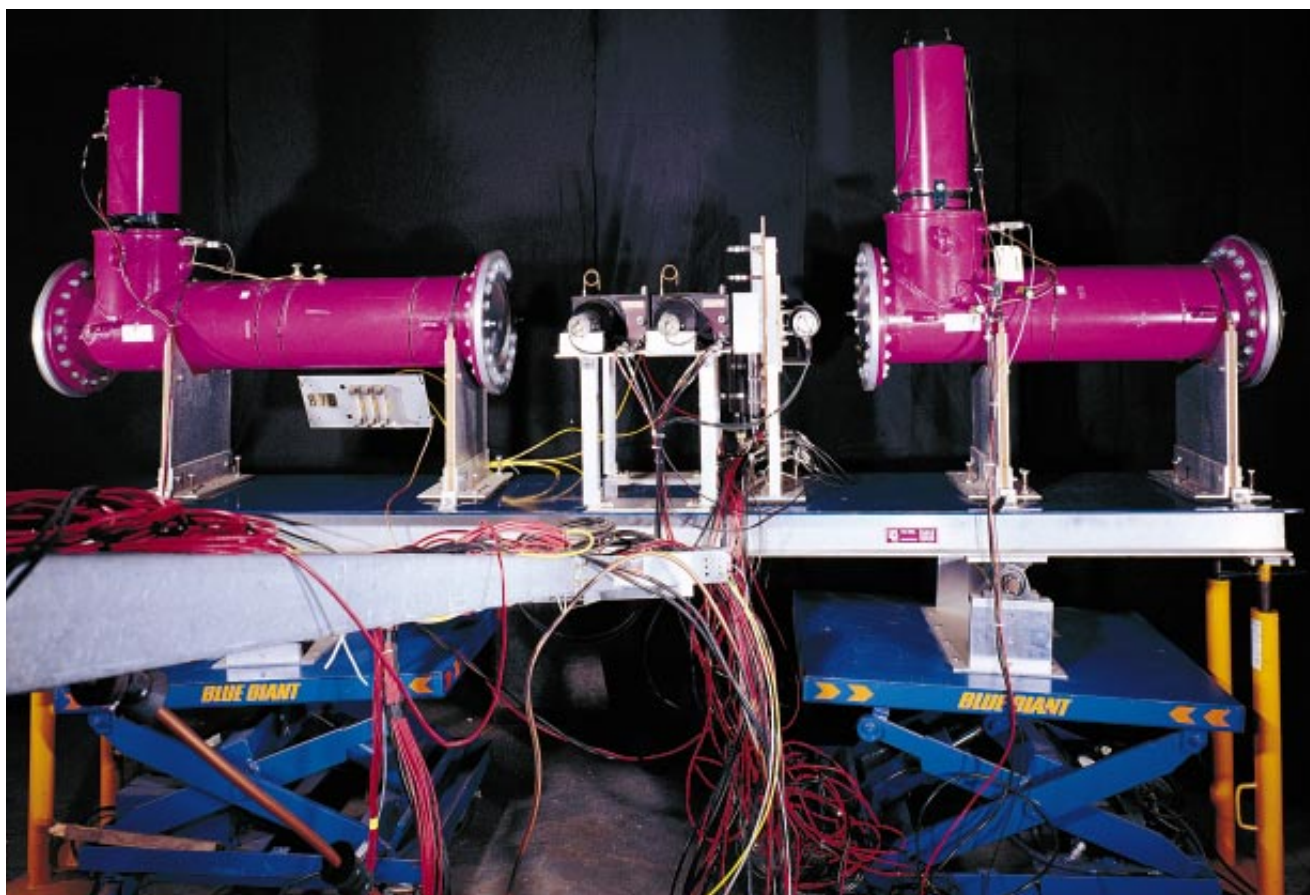


so by asking another question. Are stable bags comprising more than three quarks possible? None has yet been detected, but theorists can think of no obvious reason to forbid the existence of such objects. What is clear is that if they exist, more than just up and down quarks must make them up. To see this, consider the deuteron—the nucleus of heavy hydrogen, whose components are a proton and a neutron, or six quarks. We know from experiments that although the proton and the neutron in a deuteron are bound together, the six quarks that constitute these particles are still distinctly grouped into two three-quark bags: the proton (uud) bag and the neutron (udd) bag. This situation would not be possible if a single bag comprising all six quarks had a lower energy than the deuteron, for if it did the deuteron's quarks would spon-

taneously regroup themselves into this state. This argument may easily be generalized to other nuclei to conclude that if multi-quark bags of more than three up and down quarks were stable, matter as we know it would not exist—and neither would we.

But what might happen if strange quarks were added to up and down quark combinations? Such strange quark matter would consist of roughly equal numbers of up, down and strange quarks clustered in a single bag. In 1971 Arnold R. Bodmer of the University of Illinois was the first investigator to consider this new form of matter. He proposed that strange multi-quark clusters, being much more compressed than ordinary nuclei, may exist as long-lived exotic forms of nuclear matter inside stars.

Sui Chin and Arthur K. Kerman of the Massachusetts Institute of Technology and, independently, Larry D. McLerran of the University of Minnesota and James D. Bjorken of Stanford University took up the question. They deduced some general arguments explaining why strange quark matter should be stable. Like the electrons orbiting an atom, the quarks in a hadronic bag occupy distinct energy levels, or quantum states. According to the Pauli exclusion principle, which is the quantum analogue of Archimedes' principle that no two bodies can occupy the same space at the same time, only one quark can occupy each quantum state. One reason for the stability of strange quark matter might be that there are no empty energy states to receive the down quarks that would result from the weak decay of strange quarks: the low-ener-



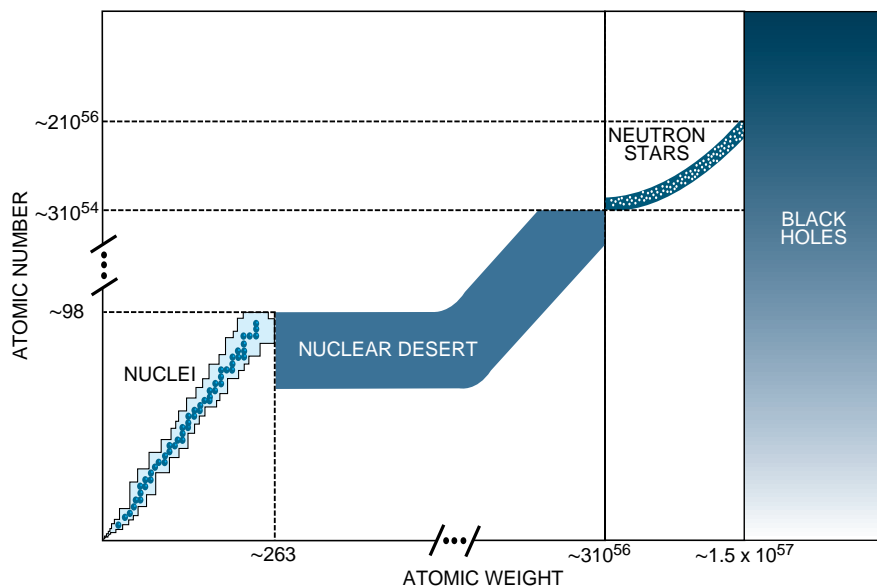


CHART OF NUCLIDES shows all known forms of stable matter. Between the heaviest atomic elements and neutron stars, which are giant nuclei, lies a vast, unpopulated nuclear desert. This void may actually be filled with strange quark matter.

gy down-quark states are already filled. This principle elucidates the stability of ordinary nuclei: a free neutron decays into a proton in about 11 minutes, but in stable nuclei, neutrons can exist virtually forever. The reason is that if the neutron were to decay, there would be no empty quantum states to receive the newly created proton. Nuclei in which there are empty energy states for the proton are radioactive and undergo beta decay.

But what could explain the ability of strange quark matter to fill in the range of sizes between the nucleus and the neutron star? Nuclear matter consists of roughly equal numbers of protons, which carry one unit of charge, and neutrons, which carry no charge at all. Electrostatic repulsion of the like-charged protons in a nucleus increases as the number of protons increases. Ultimately the electrostatic repulsion overwhelms the strong force that binds nuclei together, which is why there is a limit to the size of stable nuclei.

The situation in a quark bag that holds strange matter differs significantly. The laws of quantum mechanics dictate that, at equilibrium, the three quark flavors in the multi-quark bag share the available energy equally. The strange quark is more massive than the up or the down, so there will be slightly fewer strange quarks in a chunk of strange quark matter (mass and energy being equivalent). The electrical charge of the up quark, which is $+2/3$ that of an electron, will therefore be largely (but not completely) canceled by the sum of the $-1/3$ charges carried by each

down and strange quark. As a result, strange quark matter should carry only a very slight positive charge and, because of the near balance between positive and negative charges, should thus be free of the size limitation that affects ordinary nuclear matter. Huge chunks of stable strange quark matter could therefore exist.

If they do, their presence could resolve a long-standing astrophysical enigma. From detailed observations of galaxies, astrophysicists have concluded that there is far more to the universe than meets the eye. The combined gravitational fields of all visible stars and luminous galactic dust are not close to being strong enough to produce the motions of the galaxies or of individual stars within them. Calculations show that the amount of missing material is enormous; at least 80 percent of all the matter in the universe is apparently cold and dark, undetectable by any radio or optical telescope.

In 1984 Edward Witten of Princeton University raised the intriguing possibility that the missing mass—that is, most of the universe—is strange quark matter. Witten's scenario begins in the very early universe, shortly after the big bang but before light nuclei began to form. The cosmos was then so hot and dense that quarks wandered freely. Witten postulates that strange quark matter formed from this quark phase within the first 10^{-6} second after the big bang. The diameter of each of these nuggets was between 10^{-7} and 10 centimeters. Between 10^{33} and 10^{42} quarks made up each nugget, and each nugget

weighed from 10^9 and 10^{18} grams. A nugget the size of a baseball would weigh over a trillion tons. But because these nuggets are so small, they would scatter very little light and would be almost impossible to observe directly.

By adapting calculations used to predict the mass of ordinary hadrons, Edward H. Farhi and Robert L. Jaffe of M.I.T. have found that chunks of strange quark matter, or strangelets, could be stable for a much larger range of sizes than Witten predicts. If Farhi and Jaffe are right, strange quark matter could fill the gigantic nuclear desert. This speculative picture cannot be ruled out by any of the known principles of physics.

The alert reader, however, might fear a potentially catastrophic consequence of the existence of strangelets lighter than ordinary nuclei: ordinary matter would decay into them. Farhi and Jaffe assure us that although this could happen, the probability is so small that it is unlikely to happen in a time span many times longer than the current age of the universe.

If strange quark matter does account for 80 percent of the mass of the universe, it seems logical that occasionally a chunk of it should fall to earth. Alvaro De Rujula of CERN, the European laboratory for particle physics near Geneva, and Sheldon L. Glashow of Harvard University have calculated the effects of such encounters. A strangelet of less than about 10^{14} quarks, they determined, could be slowed and stopped by the earth. Such encounters could take the form of unusual meteorite events, earthquakes with special signatures or peculiar particle tracks in ancient mica. Nuggets of more than 10^{23} quarks would have too much momentum to be stopped by the encounter. They would instead simply pass through the earth. The sizes predicted by Witten's scenario might not be observed at all.

Nuggets less than about 10^7 quarks in size may have broken off from larger clusters and become embedded in meteoric or crustal material, where they would behave much like typical nuclei. At the University of Mainz, Klaus Lützenkirchen and his German and Israeli colleagues have recently begun to search for small strangelets in meteorites. Lützenkirchen has devised an ingenious method of screening his meteorites for strangelets. His technique relies on the fact that strangelets are much heavier than ordinary nuclei. He fires a beam of uranium nuclei at the meteorites and looks for those that bounce directly backward, as if they

had hit a brick wall. Elementary physics can prove that this happens only when the mass of the target is greater than that of the uranium nucleus. Thus far these and other experiments have produced no evidence for the existence of stable strange matter, although they have placed some limits on the range of their masses.

Cosmological observations have also been used by several researchers to place limits on the amount of strange quark matter in the universe. If strange nuggets were formed in the big bang, they would have absorbed neutrons, thus lowering the ratio of free neutrons to protons. This effect in turn would lower the rate of production of the isotope helium 4. The rate of absorption of neutrons, and therefore the rate of helium production, is very sensitive to the total surface area of all the nuggets present. For a fixed amount of mass, surface area depends on the size and number of the particles: the total surface area of many small fragments far exceeds the surface area of a few large pieces, even though both collections have the same mass. Hence, the smaller (and more numerous) the individual nuggets, the greater the total absorption of neutrons.

K. Riisager and J. Madsen of the University of Aarhus in Denmark quanti-

fied this argument. The scientists found that the primordial quark nuggets had to be made up of more than 10^{23} quarks if their existence were to be consistent with both the calculated amount of missing dark matter and the observed abundance of light isotopes.

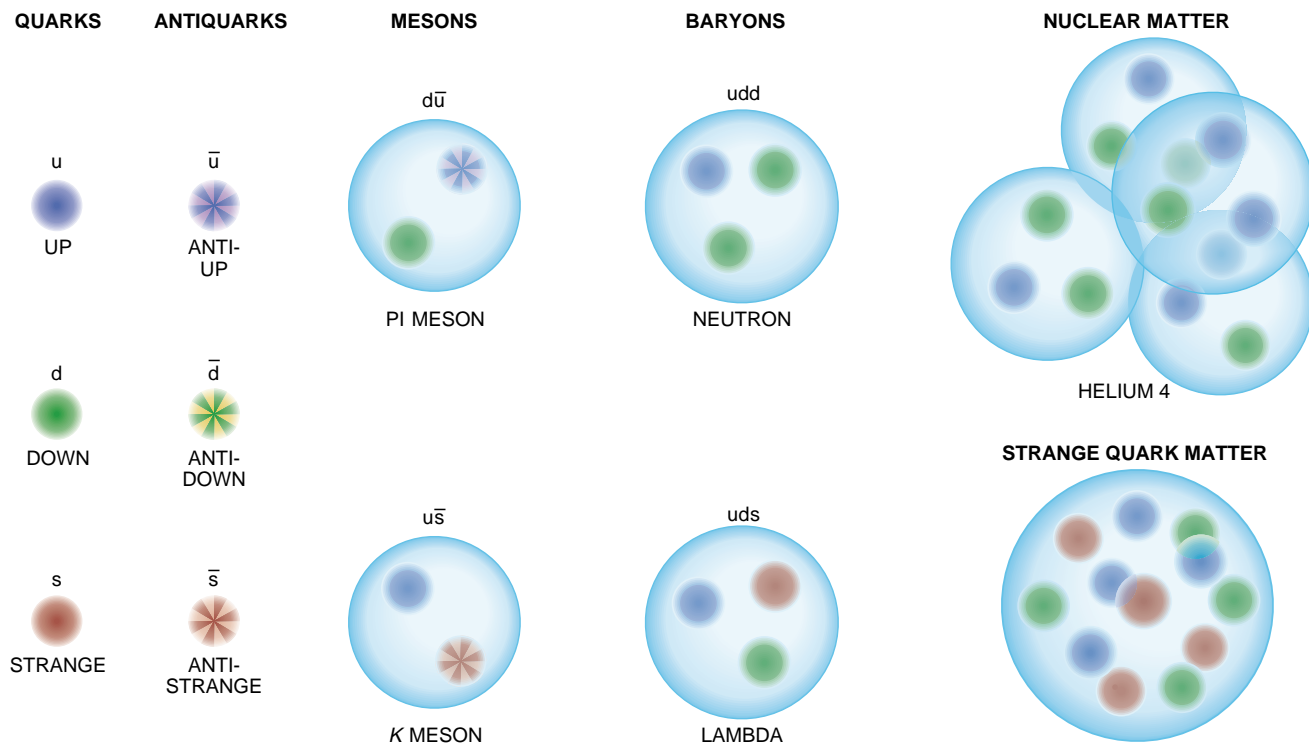
Strange matter might also be found in the superdense neutron stars that are the remnants of supernovae. A droplet of strange matter falling on a neutron star would attack it like a virus, gobbling up neutrons. The reason for this rapacity is that neutrons, being electrically neutral, do not repel the approach of the droplet, which has a small positive charge; the neutron's quarks are absorbed by the droplet. Angela V. Olinto of the University of Chicago has shown that a strange droplet could consume a neutron star, changing it from a neutron star into a strange star, in less than one minute. A strange star would be more compact than a neutron star because it is bound by intrinsic quark forces.

Like a spinning ice skater with her arms drawn in, the smaller strange star would rotate more rapidly than a neutron star, and the rate of this rotation could be detected. The observation of a half-millisecond pulsar would be strong evidence of the existence of a strange

star since ordinary neutron stars cannot spin this rapidly. Astrophysicists are eagerly seeking such objects.

In the absence of rapidly twirling strange stars, it is unlikely that strange matter can be detected by current techniques of observational astronomy. Nuclear and particle physicists have recently begun to look for more direct evidence of strange quark matter, employing powerful particle accelerators. By causing two heavy nuclei to collide head-on at the highest available energies, experimenters are now in the fortunate position of being able to simulate in the laboratory many of the conditions of the early universe. Such "little big bangs" offer a deft tool for producing exciting and unexpected arrangements of quarks at high temperatures and pressures [see "Hot Nuclear Matter," by Walter Greiner and Horst Stöcker; *SCIENTIFIC AMERICAN*, January 1985; and "The Nuclear Equation of State," by Hans Gutbrod and Horst Stöcker; *SCIENTIFIC AMERICAN*, November 1991].

Formation of the little bangs requires collisions of the heaviest nuclei at the highest attainable energies. When heavy ions, such as gold and lead, strike one another, shock waves are triggered that heat up the nuclear matter. The energy of the nuclei leads to the production of



QUARKS IN VARIOUS COMBINATIONS form all known hadronic particles. Only the lightest two, "up" and "down" quarks, are needed to make ordinary matter of the kind that accounts for the world around us and the visible universe. A third type, the "strange" quark, has so far been found only in

unstable particles. Under normal conditions, quarks behave as though they were confined in bags in which they can move freely but from which they cannot escape. Baryons consist of three quarks; mesons of a quark and an antiquark. No other combinations of quarks have yet been observed.

a fireball and the formation of a flood of exotic hadrons.

Two heavy nuclei, colliding at enormous energies, can be thought of as two drops of liquid. On collision the temperature of the liquid soars. As it does so, it undergoes a phase transition and becomes a gas composed of all kinds of hadronic particles. If the energy of the collision is high enough, the bags of the individual hadrons will rupture, freeing the quarks to roam. The nuclear matter experiences a second phase transition, becoming a free quark-gluon plasma that resembles the state of the universe immediately after the big bang. (Gluons are the particles that, under normal conditions, bind quarks together.) The plasma will comprise the up and down quarks of the original nuclei, plus equal numbers of strange quarks and antiquarks, created directly from the energy of the collision.

Just as it did in the moments that followed the big bang, the quark-gluon plasma rapidly begins to cool. The quarks condense back into bags in a process known as hadronization. Providing direct proof of this fleeting instant of the quark-gluon plasma's existence turns out to be a complicated task. Quark droplets may form during this transition from plasma to hadron gas and live long enough to be observed. The mechanism for the formation of strangelets out of a cooling quark-gluon plasma was first proposed by Han-Chao Liu and Gordon L. Shaw of the University of California at Irvine and, independently, by Peter Koch of the

University of Bremen, Horst Stöcker of the University of Frankfurt and one of us (Greiner). They hypothesized that the antistrange quarks that are found in equal number to the strange quarks in the quark-gluon plasma (strange quarks and their antimatter twins must be produced in pairs) combine with the abundant light up and down quarks of the original nuclei to form *K* mesons. Producing strange baryons, such as the lambda, from the surplus strange quarks, according to calculations by Stöcker and Greiner, is energetically more expensive than producing strangelets. This hypothesis suggests that if strange quark matter exists at low temperatures, it should condense out of a cooling mass of quark-gluon plasma.

To detect nuggets of strange quark matter, experimentalists must devise ways to separate them from the shower of normal hadronic matter. The difficulty is that they constitute a new form of matter, not a specific type of particle. Usually an investigator designs an experiment to find a particle of a single, well-defined mass. But droplets of strange quark matter can have almost any mass.

The key to detecting nuggets of strange quark matter is to take advantage of their previously mentioned small charge-to-mass ratio. For normal nuclear matter, this ratio ranges from 1:3 for the hydrogen isotope of tritium, which contains two neutrons and a proton, to 1 for the single proton of a hydrogen nucleus. Most

nuclei have roughly the same number of protons and neutrons, which gives them a charge-to-mass ratio of 1:2. In contrast, strange matter should have a charge-to-mass ratio as small as $\pm 1:10$ or $\pm 1:20$, making it easy to distinguish from ordinary matter.

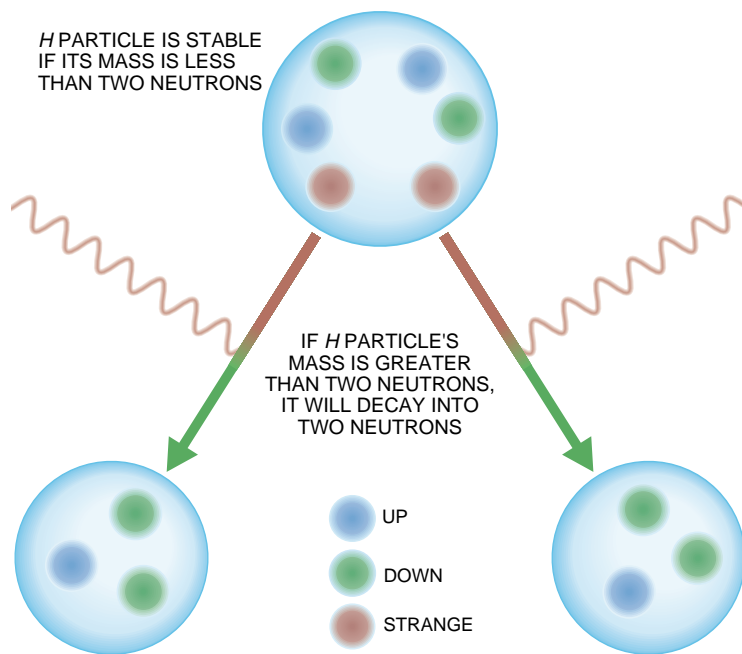
A magnetic spectrometer is the instrument of choice. In a magnetic spectrometer, beams of charged particles are deflected by a very strong magnetic field. By measuring the angle of deflection and the velocity of the particle as it enters the magnetic spectrometer, it is easy to obtain a particle's charge-to-mass ratio. Several experiments are currently under way that use this technique to search for strange quark matter.

The first highly sensitive search for strange quark matter and other particles created in high-energy nuclear collisions is now being performed by one of us (Crawford) and his colleagues from the U.S. and Japan at Brookhaven National Laboratory. In this experiment, a beam of gold nuclei, traveling at nearly the speed of light, smashes into a target made of gold foil. Between 500 and 1,000 particles are produced in each collision. The Brookhaven experiment examines only the few particles that are traveling in the direction of the beam and focused by a series of magnets.

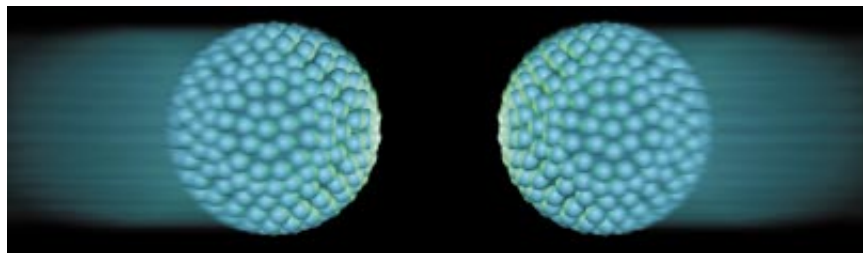
The particles first encounter the magnetic spectrometer, where the angle of deflection produced as they pass through a powerful magnetic field is measured. Next the particles' velocities are measured. The measurement is ac-

Stability of Strange Quark Matter

In 1977 Robert L. Jaffe of the Massachusetts Institute of Technology considered the possibility that particles containing more than three quarks might be stable. He started by imagining a bound state of two lambda particles, each of which is made of an up, a down and a strange quark. He called this state the *H* particle and pointed out that in order for it to be stable it would have to weigh less than two lambda particles. Otherwise, it would quickly decay into two lambdas. He also realized that the *H* particle must weigh less than two neutrons for it to be absolutely stable. If not, the two strange quarks would each decay via the weak interaction into a down quark. The resulting quarks could then form two neutrons. Unfortunately, accurately calculating the mass of the *H* particle from the Standard Model is beyond the current ability of physicists.



“LITTLE BIG BANG” is created in a particle accelerator when two heavy nuclei (*top*) collide, producing a hot plasma of quarks and gluons, the particles that bind quarks together. As the plasma cools, most of the quarks will combine to form the hadronic particles familiar to physicists. Positive K mesons carry away the anti-strange quarks, leaving the strange quarks to form strangelets.

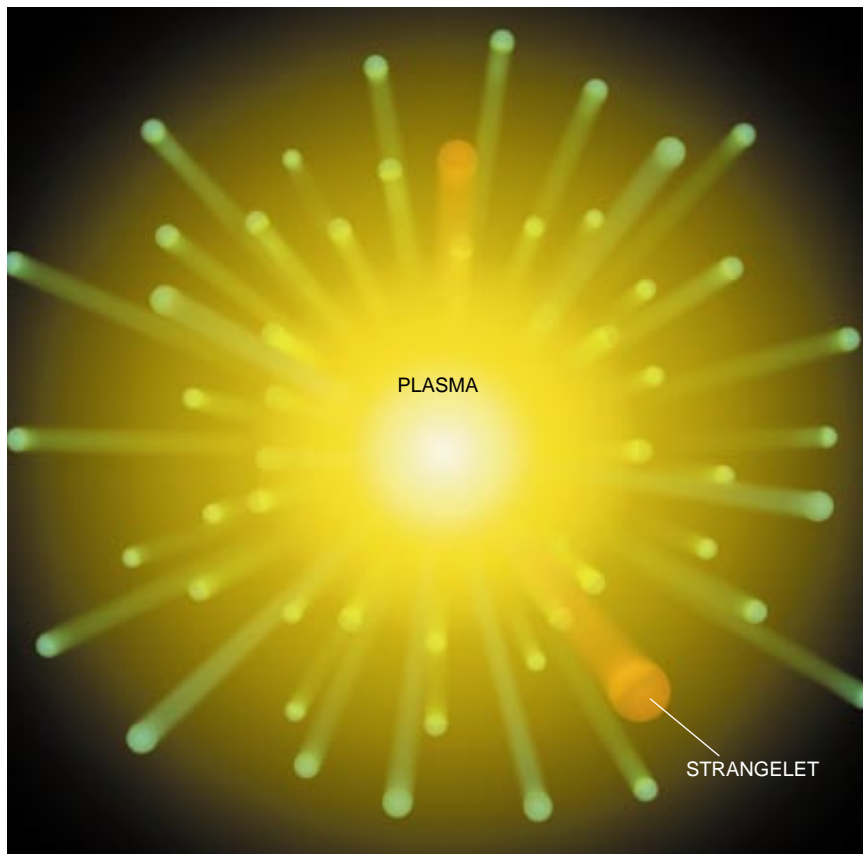


accomplished in two ways. The velocity of the slower particles is determined by observing their passage through a series of detectors known as scintillation counters, thin sheets of plastic that give off tiny flashes of light as charged particles traverse them. Velocity is calculated by measuring how long it takes the particles to pass from one detector to the other and dividing this value into the distance between the detectors. The velocity of the faster particles is measured by a Cerenkov detector. A Cerenkov counter exploits the fact that when a charged particle passes through a medium at a speed greater than the speed of light in that medium, it emits a glowing shock wave. Combining the deflection angle and the velocity gives the charge-to-mass ratio.

The spectrometer being built at Brookhaven is relatively simple. A major limitation is that this detector can see only particles emerging at a tiny angle from the beam. It is much like looking at an object through a high-power microscope. The image may be sharp, but the area viewed is tiny, and so finding a minute object in a large field becomes difficult. The Brookhaven spectrometer also has a narrow range of sensitivity. Strange matter whose charge-to-mass ratio is lower than 1:25 will not be detected.

To increase the detector's limited sensitivity, one can either lower the magnification or build a bigger detector. Both approaches are being taken by different teams of physicists searching for strange quark matter. P. Buford Price and his co-workers at the University of California at Berkeley have adopted the first approach. Their experiment is sensitive to heavy, slow particles that are deflected very little by a magnetic field.

At Brookhaven, Jack Sandweis of Yale University and his colleagues are pursuing the second method. They are constructing a gigantic nonfocusing, or open geometry, spectrometer. With no focusing magnets, their experimental apparatus is almost 30 meters long. Its detectors are eight meters wide and three meters high. Since the device is so large, many particles enter the spec-



trometer after each collision, which adds to the complexity of the operation.

CERN, whose accelerators provide much higher energy particles than those attainable at Brookhaven, is also beginning a program to look for new forms of matter. Klaus Pretzl of Bern University in Switzerland and his co-workers plan to stage collisions between lead nuclei. They will use a spectrometer similar to the one that is installed at Brookhaven. The CERN setup is, however, almost 500 meters long. This exper-

iment will take its first beam in 1994.

Inspired by theoretical computations, the search for strange quark matter is now well under way. Its detection—either on the earth, in the skies or in subatomic collisions within the world's most potent particle accelerators—would help elucidate the nature of quarks, the structure of matter and the composition of the universe. The discovery would also prove that the world is as strange a place as physicists imagine it to be.

FURTHER READING

INTRODUCTION TO HIGH-ENERGY PHYSICS. Donald H. Perkins. Addison-Wesley Publishing, 1987.
 FROM QUARKS TO THE COSMOS: TOOLS OF DISCOVERY. Leon M. Lederman and David N. Schramm. Scientific American Library, 1989.
 THE NUCLEAR EQUATION OF STATE, Part

B: QCD AND THE FORMATION OF THE QUARK-GLUON PLASMA. Edited by Walter Greiner and Horst Stöcker. Plenum Press, 1989.
 SIMULATING HOT QUARK MATTER. Jean Potvin in *American Scientist*, Vol. 79, No. 2, pages 116-129; March/April 1991.

The Toxins of Cyanobacteria

These poisons, which periodically and fatally contaminate the water supplies of wild and domestic animals, can also harm humans. But they are being coaxed into doing good

by Wayne W. Carmichael

On May 2, 1878, George Francis of Adelaide, Australia, published the first scholarly description of the potentially lethal effects produced by cyanobacteria—the microorganisms sometimes called blue-green algae or, more colloquially, pond scum. In a letter to *Nature* he noted that an alga he thought to be *Nodularia spumigena* had so proliferated in the estuary of the Murray River that it had formed a “thick scum like green oil paint, some two to six inches thick, and as thick and pasty as porridge.” This growth had rendered the water “unwholesome” for cattle and other animals that drink at the surface, bringing on a rapid and sometimes terrible death:

Symptoms—stupor and unconsciousness, falling and remaining quiet, as if asleep, unless touched, when convulsions come on, with head and neck drawn back by rigid spasm, which subsides before death. Time—sheep, from one to six or eight hours; horses, eight to twenty-four hours; dogs, four to five hours; pigs, three or four hours.

Since 1878, investigators have confirmed that *Nodularia* and many other genera of cyanobacteria include poisonous strains. Indeed, such microbes are known to account for spectacular die-offs of wild and domestic animals.

WAYNE W. CARMICHAEL is professor of aquatic biology and toxicology at Wright State University. He earned a doctorate in aquatic toxicology at the University of Alberta in Edmonton in 1974. After completing a postdoctoral appointment at Alberta, he joined Wright State as an assistant professor in 1976. Carmichael is currently engaged in determining the distribution of toxic cyanobacteria around the world, exploring methods for detecting toxins in water supplies and applying the methods of biotechnology to the study of bioactive molecules in cyanobacteria and algae.

In the midwestern U.S., for instance, migrating ducks and geese have perished by the thousands after consuming water contaminated by toxic cyanobacteria. In recent years, workers have identified the chemical structure of many cyanobacterial toxins and have also begun to decipher the steps by which the poisons can lead to suffering and death.

Such research is exciting interest today, in part because of worry over public health. No confirmed human death has yet been attributed to the poisons. But runoff from detergents and fertilizers is altering the chemistry of many municipal water supplies and swimming areas, increasing the concentration of nitrogen and phosphorus. These nutrients promote reproduction by dangerous cyanobacteria and thus foster formation of the dense growths, known as waterblooms, described by Francis. As cyanobacterial waterblooms become more common in reservoirs, rivers, lakes and ponds, the likelihood grows that people will be exposed to increased doses of toxins. (Water-treatment processes only partially filter out cyanobacteria and dilute their toxins.) The risk of animal die-offs grows as well.

The possibility of increased exposure has become particularly disturbing because some evidence suggests that certain cyanobacterial toxins might contribute to the development of cancer. Knowledge of the chemical structure and activity of the toxins should help scientists to devise more sensitive ways to measure the compounds in water and to develop antidotes to lethal doses. Improved understanding of how these chemicals function should also facilitate efforts to determine the long-term effects of exposure to nonlethal doses.

Research into the structure and activity of the toxins is sparking interest on other grounds as well. They and their derivatives are being considered as potential medicines for Alzheimer's disease and other disorders. The substanc-

es already serve as invaluable tools for exploration of questions in cell biology.

As worrisome and wonderful as the toxins are, other aspects of cyanobacteria are perhaps more familiar to many people. For example, textbooks often feature these bacteria as nitrogen fixers. The filamentous species (which consist of individual cells joined end to end, like beads on a string) convert atmospheric nitrogen into forms that plants and animals can use in their own life processes. In this way, they fertilize agricultural land throughout the world, most notably rice paddies, where they are often added to the soil.

Cyanobacteria are known, too, for the critical insights they have provided into the origins of life and into the origins of organelles in the cells of higher organisms. The fossil record shows that cyanobacteria already existed 3.3 to 3.5 billion years ago. Because they were the first organisms able to carry out oxygenic photosynthesis, and thus to convert carbon dioxide into oxygen, they undoubtedly played a major part in the oxygenation of the air [see “The Blue-Green Algae,” by Patrick Echlin; *SCIENTIFIC AMERICAN*, June 1966]. Over time, their exertions probably helped to create the conditions needed for the emergence of aerobic organisms. At some point, theorists suggest, certain of the photosynthesizers were taken up permanently by other microbes. Eventually these cyanobacteria lost the ability to function independently and became chloroplasts: the bodies responsible for photosynthesis in plants.

It was the toxins, however, that sparked my own curiosity about cyanobacteria. That was back in the late 1960s, when I was an undergraduate majoring in botany at Oregon State University. At the time, I had the young student's usual fascination with the microscope and things microscopic. I was also intrigued by the question of how toxins—naturally produced poisons—

damage the body. In biological circles, toxins are among the compounds referred to as secondary metabolites because they are produced by living organisms but are not known to be critically important to everyday survival.

I decided to pursue both of my interests by looking into the production and action of poisons made by cyanobacteria. In 1970 I therefore became a graduate student of Paul R. Gorham at the University of Alberta in Edmonton. Gorham was one of the first scientists to study the properties of toxic cyanobacteria and had been doing so since the 1950s. Researchers in South Africa, Australia and the U.S. were carrying out related investigations, but Gorham and his colleagues had already laid much groundwork for the kinds of studies I hoped to undertake.

When I joined Gorham's group, cyanobacteria were typically referred to as blue-green algae because of the turquoise coloring of most blooms and the similarity between the microbes and true algae (both carry out photosynthesis). But Roger Y. Stanier, then at the University of California at Berkeley, was beginning to reveal the "algae" part of the name to be a misnomer.

After the electron microscope was introduced in 1950, work by Stanier and others established that two radically different types of cells exist in the contemporary world. Prokaryotic varieties—those bearing the characteristics of bacteria—have no membrane enveloping their nuclear material and usually lack membrane-bound bodies in their interior. All other cells, including those of algae and more complex plants, are eukaryotic: they contain a definite nuclear membrane and have mitochondria as well as other organelles. Stanier's subsequent examinations of cyanobacteria prompted him to note in 1971 that "these organisms are not algae; their taxonomic association with eukaryotic groups is an anachronism.... Blue-green algae can now be recognized as a major group of bacteria."

Gorham's work, and later mine, extended the research begun when cyanobacteria were still thought to be al-

gae. By the 1940s reports implicating the microorganisms in the poisoning of wild and domestic animals had accumulated from many parts of the world. The animals died after drinking from ponds or other waters partly covered by slimy carpets of what seemed to be algae, often in the dog days of late summer and early fall, when the temperature is high and the air is relatively still. Yet no firm link between specific genera of cyanobacteria and animal deaths had yet been established.

Theodore A. Olson, a microbiologist

at the University of Minnesota, made that connection in the course of studies he carried out between 1948 and 1950. Olson collected samples of waterblooms in his state and determined that they contained copious amounts of species from the cyanobacterial genera *Microcystis* and *Anabaena* (common groups of planktonic cyanobacteria). By feeding cyanobacteria from those blooms to laboratory animals, he was able to demonstrate that certain waterdwelling forms can indeed be poisonous to animals.



POND IN BEIJING has been contaminated by an overgrowth, or waterbloom, of toxic cyanobacteria (*green scum*). These bacteria, flourishing in the Grandview Garden Park, are members of the widespread genus *Microcystis*, many species of which produce potent liver toxins. The toxins have killed animals, and the consumption of low doses in drinking water is suspected of contributing to a high rate of human liver cancer in certain parts of China.



MASS OF CYANOBACTERIA close to the shore of Balgavies Loch, near Dundee, Scotland, has the typical appearance of a waterbloom seen at short range: it resembles a thick pool of green oil paint. This bloom occurred in 1981 and was found to consist of species in the genus *Microcystis*.

This finding, in turn, raised new questions. Why, for example, were animals poisoned most often during the dog days of summer and fall? The answer now seems to be that cyanobacteria grow remarkably well and form waterblooms when four conditions converge: the wind is quiet or mild, and the water is a balmy temperature (15 to 30 degrees Celsius), is neutral to alkaline (having a pH of 6 to 9) and harbors an abundance of the nutrients nitrogen and phosphorus. Under such circumstances, cyanobacterial populations grow more successfully than do those of true algae. (True algae can also form waterblooms, but blooms in nutrient-rich water usually consist of toxic cyanobacteria.)

The cyanobacterial blooms by themselves probably would not harm animals if the microbes clustered far from shore. But cyanobacteria move up and down within the water to obtain light for photosynthesis and, in the process, often float to the surface. There, currents and any winds that arise can push the bacteria toward the land, causing poison-filled cells to accumulate in a thick layer near the leeward shore. Animals drinking such concentrated scum can readily consume a fatal dose.

Because the cells release toxins only when they themselves die or become old and leaky, animals usually have to

ingest whole cells to be affected. They can, however, take in a fatal dose of toxins from cell-free water if someone has treated the water with a substance, such as copper sulfate, designed to break up waterblooms. The amount of cyanobacteria-tainted water needed to kill an animal depends on such factors as the type and amount of poison produced by the cells, the concentration of the cells, as well as the species, size, sex and age of the animal. Typically, though, the required volume ranges from a few millimeters (ounces) to several liters (a few gallons). Apparently, thirsty animals are often undeterred by the foul smell and taste of contaminated water.

The early demonstration that cyanobacterial toxins were responsible for animal kills in Minnesota also raised the questions that Gorham took up in the 1950s—namely, what is the chemical nature and modus operandi of the toxins? To find answers, he first had to develop methods for maintaining cultures of toxic cyanobacteria in the laboratory. In the 1950s and 1960s Gorham and his colleagues, then at the National Research Council in Ottawa, succeeded in establishing cultures for two of the most toxic cyanobacteria: *Anabaena flos-aquae* and *Microcystis aeruginosa*. With such cultures in

hand, they were able to isolate poisons produced by the cells and identify their chemical makeup. A knowledge of chemical structure offers clues to how a molecule functions.

In 1972, soon after I arrived in Gorham's laboratory, Carol S. Huber and Oliver E. Edwards, working in Edwards's laboratory at the National Research Council, determined the chemical structure of a cyanobacterial toxin for the first time. Derived from *A. flos-aquae*, and named anatoxin-a, it turned out to be an alkaloid—one of thousands of nitrogen-rich compounds that have potent biological, usually neurological, effects. So far species from seven of 12 cyanobacterial genera involved in animal deaths have been cultured. Interestingly, none of the 12 genera grow attached to rocks or vegetation; all are planktonic, floating in water as single cells or filaments. Most produce more than one type of toxin.

The toxins that have been studied intensively to date belong to one of two groups, defined by the symptoms they have produced in animals. Some, such as anatoxin-a, are neurotoxins. They interfere with the functioning of the nervous system and often cause death within minutes, by leading to paralysis of the respiratory muscles.

Other cyanobacterial poisons, such as those produced by Francis's *N. spumigena*, are hepatotoxins. They damage the liver and kill animals by causing blood to pool in the liver. This pooling can lead to fatal circulatory shock within a few hours, or, by interfering with normal liver function, it can lead over several days to death by liver failure.

Four neurotoxins have been studied in detail. Of these, anatoxin-a and anatoxin-a(s) seem unique to cyanobacteria. The other two—saxitoxin and neosaxitoxin—arise in certain marine algae as well. I had the good fortune of being able to explore the activity of anatoxin-a soon after its structure was deciphered. This compound, made by various strains of the freshwater genera *Anabaena* and *Oscillatoria*, mimics the neurotransmitter acetylcholine.

When acetylcholine is released by neurons (nerve cells) that impinge on muscle cells, it binds to receptor molecules containing both a neurotransmitter binding site and an ion channel that spans the cell membrane. As acetylcholine attaches to the receptors, the channel opens, triggering the ionic movement that induces muscle cells to contract. Soon after, the channel closes, and the receptors ready themselves to respond to new signals. Meanwhile an enzyme called acetylcholinesterase de-

grades the acetylcholine, thereby preventing overstimulation of the muscle cells.

Anatoxin-a is deadly because it cannot be degraded by acetylcholinesterase or by any other enzyme in eukaryotic cells. Consequently, it remains available to overstimulate muscle. It can induce muscle twitching and cramping, followed by fatigue and paralysis. If respiratory muscles are affected, the animal may suffer convulsions (from lack of oxygen to the brain) and die of suffocation. Unfortunately, no one has succeeded in producing an antidote to anatoxin-a. Hence, the only practical way for farmers or other concerned individuals to prevent deaths is to recognize that a toxic waterbloom may be developing and to find an alternative water supply for the animals until the bloom is eliminated.

For animals, anatoxin-a is an anathema, but for scientists it is a blessing. As a mimic of acetylcholine, anatoxin-a makes a fine research tool. For example, because it resists breakdown by acetylcholinesterase, the toxin and its derivatives can be used in place of acetylcholine in experiments examining how acetylcholine binds to and influences the activity of acetylcholine receptors (especially the so-called nicotinic acetylcholine receptors in the peripheral and central nervous system).

Edson X. Albuquerque and his colleagues at the University of Maryland School of Medicine are looking at anatoxin-a in other ways as well. The researchers are in the early stages of exploring the intriguing possibility that a modified version might one day be administered to slow the mental degeneration of Alzheimer's disease. In many patients, such deterioration results in part from destruction of neurons that produce acetylcholine. Acetylcholine itself cannot be administered to replace the lost neurotransmitter because it disappears too quickly. But a version of anatoxin-a that has been modified to reduce its toxicity might work in its place. Derivatives of anatoxin-a could also conceivably prove useful for other disorders in which acetylcholine is deficient or is prevented from acting effectively, such as myasthenia gravis (a degenerative disorder that causes muscle weakness).

The other neurotoxin unique to cyanobacteria, anatoxin-a(s), is made by strains of *Anabaena*. It produces many of the same symptoms as anatoxin-a—which is how it came to have such a similar name. The letter “s” was appended because anatoxin-a(s) seemed to be a variant of anatox-

in-a that caused vertebrates to salivate excessively. Recently, however, my students and I at Wright State University, together with Shigeki Matsunaga and Richard E. Moore of the University of Hawaii, have shown that anatoxin-a(s) differs chemically from anatoxin-a and elicits symptoms by other means.

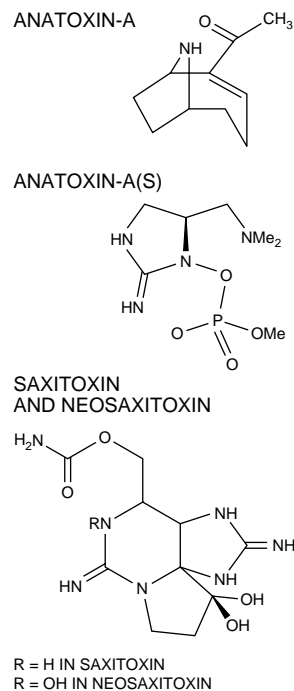
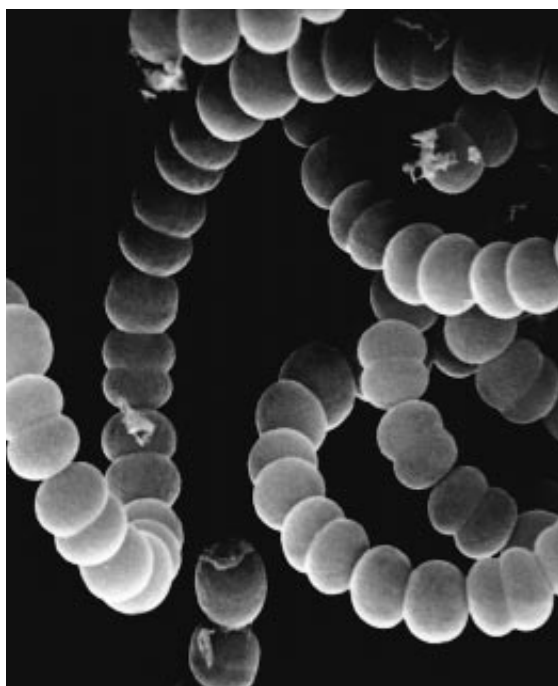
Anatoxin-a(s) is a naturally occurring organic phosphate that functions much like synthetic organophosphate insecticides, such as parathion and malathion. To my knowledge, it is the only natural organophosphate yet discovered. Even though its structure differs from that of the synthetic compounds, its killing power, like theirs, stems from its ability to inhibit acetylcholinesterase. By impeding acetylcholinesterase from degrading acetylcholine, it ensures that acetylcholine remains continuously available to stimulate—and overstimulate—muscle cells.

As a structurally novel organophosphate, anatoxin-a(s) could in theory form the basis for new pesticides. Synthetic organophosphates are widely used because they are more toxic to insects than to humans. They are, however, under some fire. Soluble in lipids (fats), they tend to accumulate in cell membranes and other lipid-rich parts of humans and other vertebrates. Anatoxin-a(s), in contrast, is more soluble in water and, hence, more biodegrad-

able. So it could be safer. On the other hand, it might also be less able to cross the lipid-rich cuticles, or exoskeletons, of insects. By tinkering with the structure of anatoxin-a(s), investigators might be able to design a compound that would minimize accumulation in tissues of vertebrates but continue to kill agricultural pests.

As is true of anatoxin-a and anatoxin-a(s), the neurotoxins saxitoxin and neosaxitoxin disrupt communication between neurons and muscle cells. But they do so by preventing acetylcholine from being released by neurons. In order to secrete acetylcholine or other neurotransmitters, neurons must first generate an electrical impulse. Then the impulse must propagate along the length of a projection called an axon—an activity that depends on the flow of sodium and potassium ions across channels in the axonal membrane. When the impulse reaches an axon terminal, the terminal releases stores of acetylcholine. Saxitoxin and neosaxitoxin block the inward flow of sodium ions across the membrane channels; in so doing, they snuff out any impulses and forestall the secretion of acetylcholine.

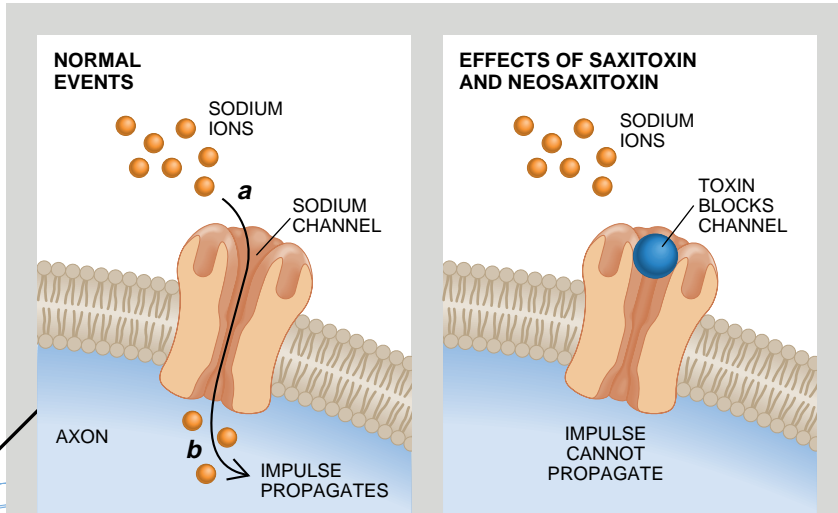
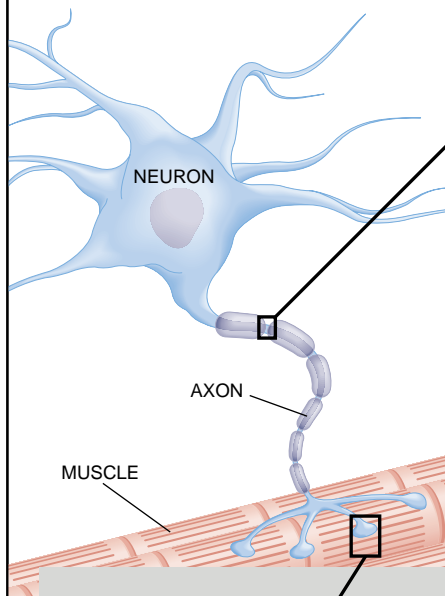
Although saxitoxin and neosaxitoxin occur in some strains of the cyanobacterial genera *Anabaena* and *Aphanizomenon*, these poisons are actually better known as products of dinoflagel-



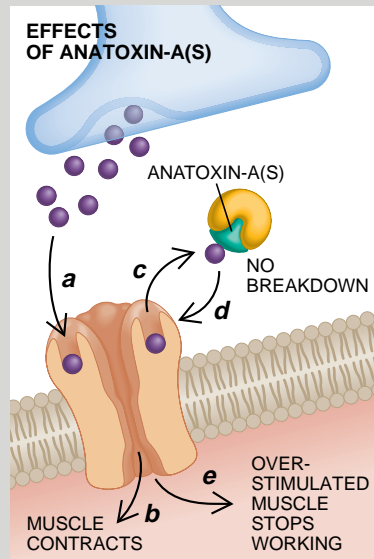
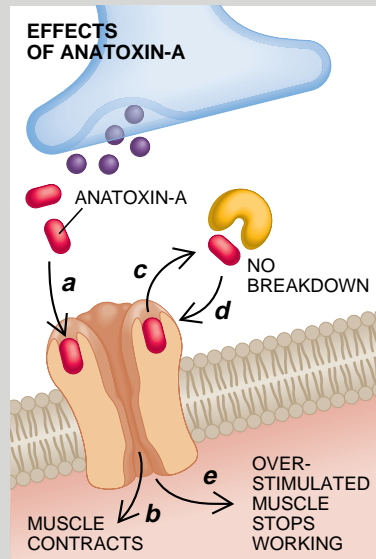
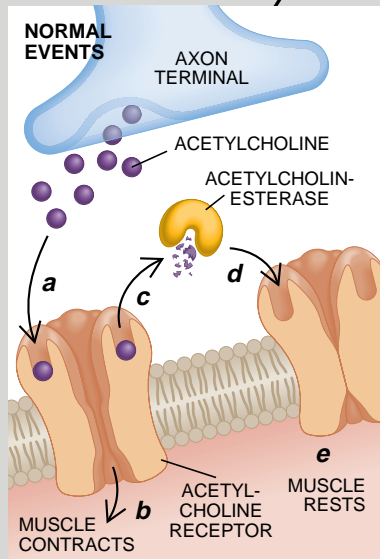
BEADS ON A STRING (micrograph) are actually cells of the cyanobacterium *Anabaena flos-aquae*, magnified some 2,500 times. *A. flos-aquae* is a major producer of neurotoxins, poisons that interfere with the functioning of the nervous system. The strain shown here was responsible for the death of hogs in Griggsville, Ill. The chemical structures at the right represent toxins made by strains of *Anabaena*; all except anatoxin-a(s) also occur in other cyanobacteria.

How Neurotoxins Kill

Neurotoxins produced by cyanobacteria can disrupt normal signaling between neurons and muscles in several ways. All of them lead to death by causing paralysis of respiratory muscles, followed by suffocation.



Saxitoxin and neosaxitoxin silence the neurons that act on muscle cells. Sodium ions (gold) must flow into neurons (a at left) in order for the neurons to relay impulses (b) to other cells. Saxitoxin and neosaxitoxin (blue sphere at right) halt impulse propagation by preventing the ions from passing into the neurons. When the nerve cells are thus quieted, muscle cells receive no stimulation and become paralyzed.



Anatoxin-a and anatoxin-a(s) (center and right panels) overexcite muscle cells by disrupting the functioning of the neurotransmitter acetylcholine. Normally, acetylcholine molecules (purple) bind to acetylcholine receptors on muscle cells (a in left panel), thereby inducing the cells to contract (b). Then the enzyme acetylcholinesterase (yellow) degrades acetylcholine (c), allowing its receptors and hence the muscle cells to return to their resting state (d and e). Anatoxin-a (red in center panel) is a mimic of acetylcholine. It, too, binds to acetylcholine receptors (a), triggering con-

traction (b), but it cannot be degraded by acetylcholinesterase (c). Consequently, it continues to act on muscle cells (d). The cells then become so exhausted from contracting that they stop operating (e). Anatoxin-a(s) (green in right panel) acts more indirectly. It allows acetylcholine to bind to its receptors and induce contraction as usual (a and b), but it blocks acetylcholinesterase from degrading acetylcholine (c). As a result, the neurotransmitter persists and overstimulates respiratory muscles (d), which once again eventually become too fatigued to operate (e).

lates—the marine algae that have caused “red tides” (red waterblooms) in several coastal areas of the world. These red tides have led to repeated outbreaks of paralytic shellfish poisoning and to the closure of shellfish beds in those areas.

The discovery of saxitoxin and neosaxitoxin in cyanobacteria added few new ideas for drugs or insecticides or for ways to solve problems in cell biology, since the chemicals were already known entities. The finding did pose a fascinating riddle, however. What would cause freshwater cyanobacteria to produce the same chemicals made by marine eukaryotes? Did these disparate groups evolve the same pathways of synthesis independently, or did they perhaps share a common ancestor?

That particular puzzle remains unsolved, but the realization that cyanobacteria produce saxitoxin and neosaxitoxin has made it possible to unravel another scientific knot. For years, the biosynthetic pathway for production of the toxins was unknown because dinoflagellates were difficult to cultivate in the laboratory. Studies of more readily grown species of *Aphanizomenon* allowed Yuzuru Shimizu and his students at the University of Rhode Island to work out the pathway in 1984.

Cyanobacterial neurotoxins, then, are both deadly and potentially valuable, but they are not as ubiquitous as the other major class of cyanobacterial poisons: the hepatotoxins. Whereas neurotoxins have been blamed for kills mainly in North America (with some in Great Britain, Australia and Scandinavia), hepatotoxins have been implicated in incidents occurring in virtually every corner of the earth. For this reason, great excitement ensued in the early 1980s, when a group headed by Dawie P. Botes, then at the Council for Scientific and Industrial Research in Pretoria, determined the chemical structure of a liver toxin. Such toxins were long known to be peptides (small chains of amino acids), but the technological advances needed for determining the precise structures of the toxins did not occur until the 1970s.

Soon after Botes established the chemical identity of the first few hepatotoxins, my laboratory and others confirmed his results and began identifying the chemical makeup of other hepatotoxins. Extensive structural analyses, mainly in the laboratory of Kenneth L. Rinehart of the University of Illinois, have now established that the liver toxins form a family of at least 53 related cyclic, or ringed, peptides. Those consisting of seven amino acids are called

microcystins; those consisting of five amino acids are called nodularins. The names reflect the fact that the toxins were originally isolated from members of the genera *Microcystis* and *Nodularia*.

Research into the hepatotoxins—much of which is carried out at other laboratories with toxins supplied by my group—is directed primarily at understanding how the compounds affect the body. Investigators know that the peptides cause hepatocytes, the functional cells of the liver, to shrink. In consequence, the cells, which are normally packed tightly together, separate. When the cells separate, other cells forming the so-called sinusoidal capillaries of the liver also separate [see illustration on page 86]. Then the blood carried by the vessels seeps into liver tissue and accumulates there, leading to local tissue damage and, often, to shock.

Other details of the poisoning process are only now becoming clear. For instance, scientists have wondered why the toxins act most powerfully on the liver. The answer probably is that they are moved into hepatocytes by the transport system, found only in hepatocytes, that carries bile salts into the cells.

Maria T. C. Runnegar of the University of Southern California and Ian R. Falconer of the University of Adelaide in Australia have taken the lead in addressing the related problem of how the toxins deform hepatocytes. They, and more recently Val R. Beasley of the University of Illinois and John E. Eriksson of the Finland-Swedish University of Åbo, have found that the poisons distort liver cells by acting on the cytoskeleton: the gridwork of protein strands that, among other functions, gives shape to cells.

The cytoskeletal components most affected by the toxins are the protein polymers known as intermediate filaments and microfilaments. Subunits are continually added to and lost from the intermediate filaments, and the protein strands forming the microfilaments continually associate and dissociate. The net sizes of the intermediate filaments and of the microfilaments change little over time, however. Microcystins and nodularins seem to tilt the balance toward subunit loss and dissociation. The intermediate filaments apparently undergo change first, followed by the microfilaments. As the cytoskeleton shrinks, the fingerlike projections through which hepatocytes interact with neighboring cells withdraw, breaking the cell's contact with other hepatocytes and with sinusoidal capillaries.

Still more recent work in many laboratories offers some insight into how

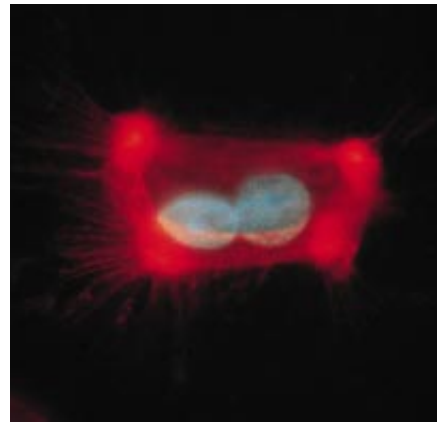
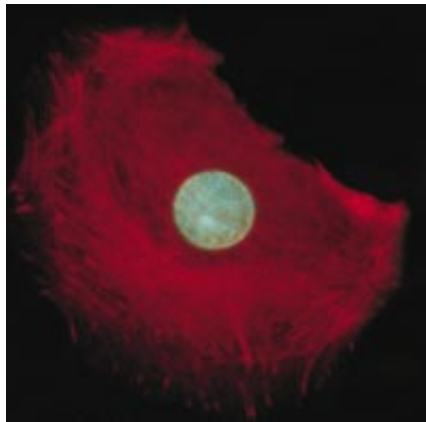
the toxins manage to disrupt cytoskeletal components. In studies of microcystins and nodularins, researchers have found that the toxins are potent inhibitors of enzymes known as protein phosphatases. These enzymes work in concert with other enzymes—protein kinases—to regulate the number of phosphate groups on proteins. The kinases add phosphate groups, and the phosphatases remove them.

Such phosphorylation and dephosphorylation reactions have long been known to influence the structure and function of intermediate filaments and microfilaments. It seems, therefore, that the toxins disrupt the fibers by upsetting the normal regulatory balance between phosphorylation and dephosphorylation. More specifically, it is thought that the unchecked activity of the kinases and the resulting excessive phosphorylation of the intermediate filaments and the microfilaments (or of proteins that act on them) increase the rate of subunit loss and dissociation.

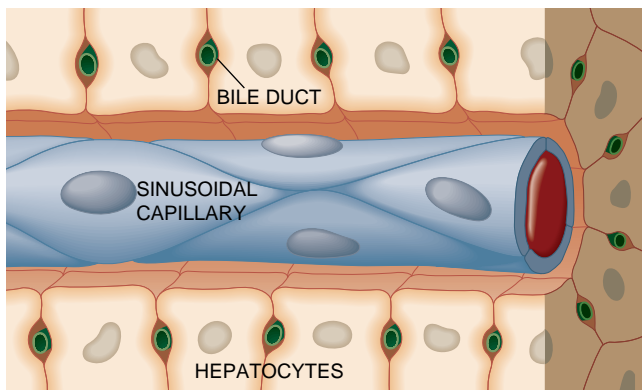
The revelation that cyanobacterial hepatotoxins can inhibit protein phosphatases has raised the disturbing possibility that human exposure to nonlethal doses might contribute to the development of cancer. Beyond influencing the structure and function of cytoskeletal fibers, protein kinases and protein phosphatases play a major part in regulating cell division. Protein kinases, which themselves are regulated by various proteins, promote movement of cells through the cell division cycle. Protein phosphatases help to check cell division by quieting the activity of the regulators. The toxins, by inhibiting the phosphatases, probably give the upper hand to the proteins that activate kinases; they may thus help release the normal brakes on cell proliferation.

Studies by Hirota Fujiki and his colleagues at the Saitama Cancer Center in Japan have now shown in cultured cells and in whole animals that microcystins and nodularins can indeed hasten tumor development. These toxins do not seem to initiate a cell's progression toward becoming cancerous; however, once something else has triggered early changes, the hepatotoxins promote development of further carcinogenic alterations. My group in Ohio and our colleagues at the Academy of Sciences in Wuhan, China, and at Shanghai Medical University are attempting to find out whether such activity might contribute to malignancy in humans. To do so, we are carrying out a long-term study of people in China who are exposed repeatedly to microcystins in

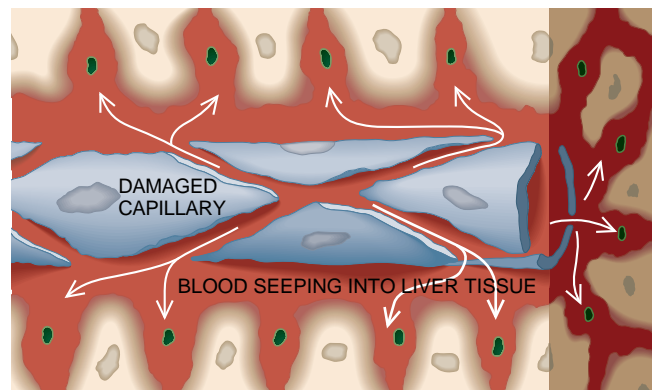
MICROFILAMENTS (*red threads in micrographs*), structural components of cells, are usually quite long, as in the rat hepatocyte at the left. But after exposure to microcystins (*right*), microfilaments collapse toward the nucleus (*blue*). (This cell, like many healthy hepatocytes, happens to have two nuclei.) Such collapse helps to shrink hepatocytes—which normally touch one another and touch sinusoidal capillaries (*left drawing*). Then the shrunken cells separate from one another and from the sinusoids (*right drawing*). The cells of the sinusoids separate as well, causing blood to spill into liver tissue. This bleeding can lead swiftly to death.



NORMAL LIVER



LIVER AFTER TOXINS ACT



It is possible, though, that the protective effect is incidental. The toxins may once have had some critical function that they have since lost. This likelihood is suggested by the fact that microcystins and nodularins act on the protein phosphatases that regulate the proliferation of eukaryotic cells. The hepatotoxins do not now seem to participate in cell function and cell division in cyanobacteria, but they may have played such a role early in the evolution of these organisms (and of other microbes as well).

Regardless of their intended purpose, the toxicity of many chemicals produced by cyanobacteria is undeniable. For this reason, I am becoming increasingly worried by a modern fad: the eating of cyanobacteria from the genus *Spirulina* as a health food. Certain tribes in Chad and many peoples in Mexico have consumed two closely related species of *Spirulina* for hundreds of years. When world health officials and scientists began looking for new high-protein food sources in the mid-1960s, many of them turned to *Spirulina* because of its high protein content. Beginning in the late 1970s certain producers and distributors of *Spirulina* began promoting it throughout large parts of the U.S., Canada and Europe as a nutritious food for humans. It has also been marketed as a diet pill, because anecdotal reports, as yet unconfirmed, indicated that a few grams taken before meals dulled the appetite.

My worry has recently intensified because the popularity of *Spirulina* has led to the production and marketing of such cyanobacteria as *Anabaena* and *Aphanizomenon*—genera that contain highly toxic strains. Some promotional material for cyanobacteria-containing products even claims that the items being sold can moderate some disease symptoms, including those of debilitating neuromuscular disorders. Yet this literature does not provide a listing of all microbial species in the marketed products, nor does it indicate that anyone is monitoring the products to ensure they are pure and nontoxic. Because cyanobacteria are often collected simply from the surface of an open body of water and because neither sellers nor buyers can distinguish toxic from nontoxic strains without applying sophisticated biochemical tests, the

safety of these items is questionable. All told, the cyanobacteria constitute a small taxonomic group, containing perhaps 500 to 1,500 species. But their power to harm and to help animals and humankind is great. Investigated and exploited responsibly, they can provide valuable tools for basic research in the life sciences and may one day participate in the treatment of disease.

FURTHER READING

METHODS IN ENZYMOLOGY, Vol. 167: CYANOBACTERIA. Edited by Lester Pack-er et al. Academic Press, 1988.

TOXIC BLUE-GREEN ALGAE: A REPORT BY THE NATIONAL RIVERS AUTHORITY. M. J. Pearson et al. National Rivers Authority, London, September 1990.

A STATUS REPORT ON PLANKTONIC CYANOBACTERIA (BLUE-GREEN ALGAE) AND THEIR TOXINS. W. W. Carmichael. U.S. Environmental Protection Agency, Report EPA/600R-92/079, June 1992.

A REVIEW OF HARMFUL ALGAL BLOOMS AND THEIR APPARENT GLOBAL INCREASE. Gustav M. Hallegraeff in *Phycologia*, Vol. 32, No. 2, pages 79-99; March 1993.

DISEASES RELATED TO FRESHWATER BLUE-GREEN ALGAL TOXINS, AND CONTROL MEASURES. W. W. Carmichael and I. R. Falconer in *Algal Toxins in Seafood and Drinking Water*. Edited by I. R. Falconer. Academic Press, 1993.

Breaking Intractability

Problems that would otherwise be impossible to solve can now be computed, as long as one settles for what happens on the average

by Joseph F. Traub and Henryk Woźniakowski

Although mathematicians and scientists must rank among the most rational people in the world, they will often admit to falling prey to a curse. Called the curse of dimension, it is one many people experi-

ence in some form. For example, a family's decision about whether to refinance their mortgage with a 15- or 30-year loan can be extremely difficult to make, because the choice depends on an interplay of monthly expenses, income, future tax and interest rates and other uncertainties. In science, the problems are more esoteric and arguably much harder to cope with. In the computer-aided design of pharmaceuticals, for instance, one might need to know how tightly a drug candidate will bind to a biological receptor. Assuming a typical number of 8,000 atoms in the drug, the biological receptor and the solvent, then because of the three spatial variables needed to describe the position of each atom, the calculation involves 24,000 variables. Simply put, the more variables, or dimensions, there are to consider, the harder it is to accomplish a task. For many problems, the difficulty grows exponentially with the number of variables.

The curse of dimension can elevate tasks to a level of difficulty at which they become intractable. Even though scientists have computers at their disposal, problems can have so many variables that no future increase in computer speed will make it possible to solve them in a reasonable amount of time.

Can intractable problems be made tractable—that is, solvable in a relatively modest amount of computer time? Sometimes the answer is, happily, yes. But we must be willing to do without a guarantee of achieving a small error in our calculations. By settling for a small error most of the

time (rather than always), some kinds of multivariate problems become tractable. One of us (Woźniakowski) formally proved that such an approach works for at least two classes of mathematical problems that arise quite frequently in scientific and engineering tasks. The first is integration, a fundamental component of the calculus. The second is surface reconstruction, in which pieces of information are used to reconstruct an object, a technique that is the basis for medical imaging.

Fields other than science can benefit from ways of breaking intractability. For example, financial institutions often have to assign a value to a pool of mortgages, which is affected by mortgagees who refinance their loans. If we assume a pool of 30-year mortgages and permit refinancing monthly, then this task contains 30 years times 12 months, or 360 variables. Adding to the difficulty is that the value of the pool depends on interest rates over the next 30 years, which are of course unknown.

We shall describe the causes of intractability and discuss the techniques that sometimes allow us to break it. This issue belongs to the new field of information-based complexity, which examines the computational complexity of problems that cannot be solved exactly. We shall also speculate briefly on how information-based complexity might enable us to prove that certain scientific questions can never be answered because the necessary computing resources do not exist in the universe. If so, this condition would set limits on what is scientifically knowable.

Information-based complexity focuses on the computational difficulty of so-called continuous problems. Calculating the movement of the planets is an example. The motion is governed by a system of ordinary differential equations—that is, equations that describe the positions of the planets as a function of time. Because time can take any real value, the mathemati-



A potentially intractable problem

cal model is said to be continuous. Continuous problems are distinct from discrete problems, such as difference equations in which time takes only integer values. Difference equations appear in such analyses as the predicted number of predators in a study of predator-prey populations or the anticipated pollution levels in a lake.

In the everyday process of doing science and engineering, however, continuous mathematical formulations predominate. They include a host of problems, such as ordinary and partial differential equations, integral equations, linear and nonlinear optimization, integration and surface reconstruction. These formulations often involve a large number of variables. For example, computations in chemistry, pharmaceutical design and metallurgy often entail calculations of the spatial positions and momenta of thousands of particles.

Often the intrinsic difficulty of guaranteeing an accurate numerical solution grows exponentially with the number of variables, eventually making the problem computationally intractable. The growth is so explosive that in many cases an adequate numerical solution cannot be guaranteed for situations comprising even a modest number of variables.

To state the issue of intractability more precisely and to discuss possible cures, we will consider the example of computing the area under a curve. The process resembles the task of computing the vertical area occupied by a collection of books on a shelf. More explicitly, we will calculate the area between two bookends. Without loss of generality, we can assume the bookends rest at 0 and 1. Mathematically, this summing process is called the computation of the definite integral. (More accurately, the area is occupied by an infinite number of books, each infinitesimally thin.) The mathematical input to this problem is called the integrand, a function that describes the profile of the books on the shelf.

Calculus students learn to compute the definite integral by following a set of prescribed rules. As a result, the students arrive at the exact answer. But most integration problems that arise in practice are far more complicated, and the symbolic process learned in school cannot be carried out. Instead the integral must be approximated numerically—that is, by a computer. More exactly, one computes the integrand values at finitely many points. These integrand values result from so-called information operations. Then one combines these values to produce the answer.

Knowing only these values does not

JOSEPH F. TRAUB and HENRYK WOŹNIAKOWSKI have been collaborating since 1973. Currently the Edwin Howard Armstrong Professor of Computer Science at Columbia University, Traub headed the computer science department at Carnegie Mellon University and was founding chair of the Computer Science and Telecommunications Board of the National Academy of Sciences. In 1959 he began his pioneering research in what is today called information-based complexity and has received many honors, including election to the National Academy of Engineering. He is grateful to researchers at the Santa Fe Institute for numerous stimulating conversations concerning the limits of scientific knowledge. Woźniakowski holds two tenured appointments, one at the University of Warsaw and the other at Columbia University. He directed the department of mathematics, computer science and mechanics at the University of Warsaw and was the chairman of Solidarity there. In 1988 he received the Mazur Prize from the Polish Mathematical Society. The authors thank the National Science Foundation and the Air Force Office of Scientific Research for their support.

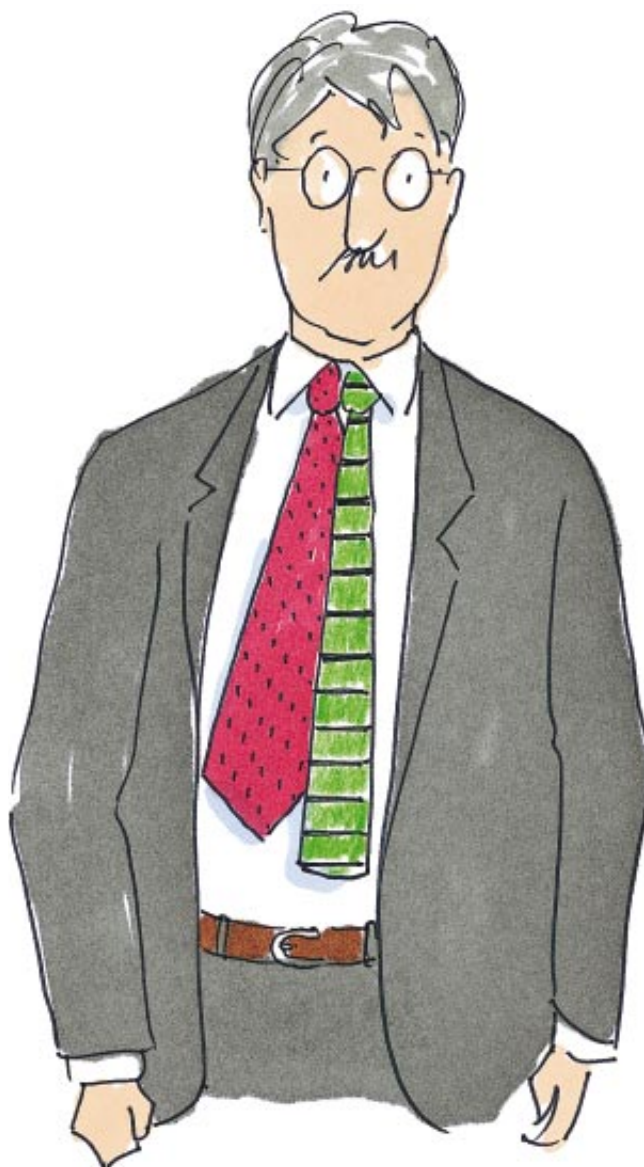
completely identify the true integrand. Because one can evaluate the integrand only at a finite number of points, the information about the integrand is partial. Therefore, the integral can, at best, only be approximated. One typically specifies the accuracy of the approximation by stating that the error of the answer falls within some error threshold. Mathematicians represent this error with the Greek letter epsilon, ϵ .

Even this goal cannot be achieved without further restriction. Knowing the integrand at, say, 0.2 and 0.5 indicates nothing about the curve between those two points. The curve can assume any shape between them and therefore enclose any area. In our bookshelf analogy, it is as if an art book has been shoved between a run of paperbacks. To guarantee an error of at most ϵ , some global knowledge of the integrand is needed. One may need to assume, for example, that the slope of the function is always less than 45 degrees—or that only paperbacks are allowed on that shelf.

In summary, an investigator trying to solve an integral must usually do it numerically on a computer. The input to the computer is the integrand values at some points. The computer produces an output that is a number approximating the integral.

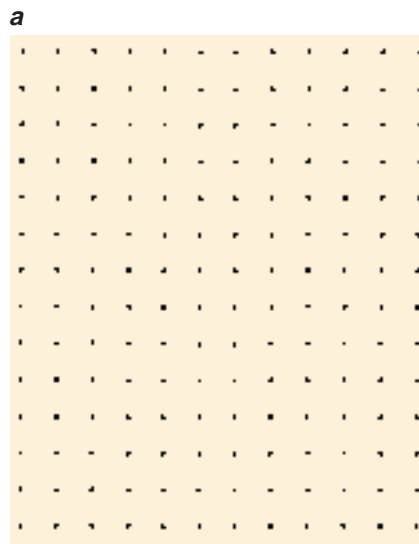
The basic concept of computational complexity can now be introduced. We want to find the intrinsic difficulty of solving the integration problem. Assume that determining integrand values and using combinatory

operations, such as addition, multiplication and comparison, each have a given cost. The cost could simply be the amount of time a computer needs to perform the operation. Then the computational complexity of this integra-



One solution to an intractable problem

SAMPLING POINTS indicate where to evaluate functions in the randomized and average-case settings. The points are plotted in two dimensions for visual clarity. The points chosen can be spaced over regular intervals such as grid points (a), or in random positions (b). Two other types, so-called Hammersley points (c) and hyperbolic-cross points (d), represent optimal places in the average-case setting.



tion problem can be defined as the minimal cost of guaranteeing that the computed answer is within an error threshold, ϵ , of the true value. The optimal information operations and the optimal combinatory algorithm are those that minimize the cost.

Theorems have shown that the computational complexity of this integration problem is on the order of the reciprocal of the error threshold ($1/\epsilon$). In other words, it is possible to choose a set of information operations and a combinatory algorithm such that the solution can be approximated at a cost of about $1/\epsilon$. It is impossible to do better. With one variable, or dimension, the problem is rather easy. The computational complexity is inversely proportional to the desired accuracy.

But if there are more dimensions to this integration problem, then the computational complexity scales exponentially with the number of variables. If d represents the number of variables, then the complexity is on the order of $(1/\epsilon)^d$ —that is, the reciprocal of the error threshold raised to a power equal to the number of variables. If one wants eight-place accuracy (down to 0.00000001) in computing an integral that has three variables, then the com-

plexity is roughly 10^{24} . In other words, it would take a trillion trillion integrand values to achieve that level of accuracy. Even if one generously assumes the existence of a sequential computer that performs 10 billion function evaluations per second, the job would take 100 trillion seconds, or more than three million years. A computer with a million processors would still take 100 million seconds, or about three years.

To discuss multivariate problems more generally, we must introduce one additional parameter, called r . This parameter represents the smoothness of the mathematical inputs. By smoothness, we mean that the inputs consist of functions that do not have any sudden or dramatic changes. (Mathematicians say that all partial derivatives of the function up to order r are bounded.) The parameter takes on nonnegative integer values; increasing values indicate more smoothness. Hence, $r=0$ represents the least amount of smoothness (technically, the integrands are only continuous—they are rather jagged but still connected as a single curve).

Numerous problems have a computational complexity that is on the order of $(1/\epsilon)^{d/r}$. For those of a more technical persuasion, multivariate integra-

tion, surface reconstruction, partial differential equations, integral equations and nonlinear optimization all have this computational complexity.

If the error threshold and the smoothness parameter are fixed, then the computational complexity depends exponentially on the number of dimensions. Hence, the problems become intractable for high dimensions. An impediment even more serious than intractability may occur: a problem may be unsolvable. A problem is unsolvable if one cannot compute even an approximation at finite cost. This is the case when the mathematical inputs are continuous but jagged. The smoothness pa-

Developing a Random Approach

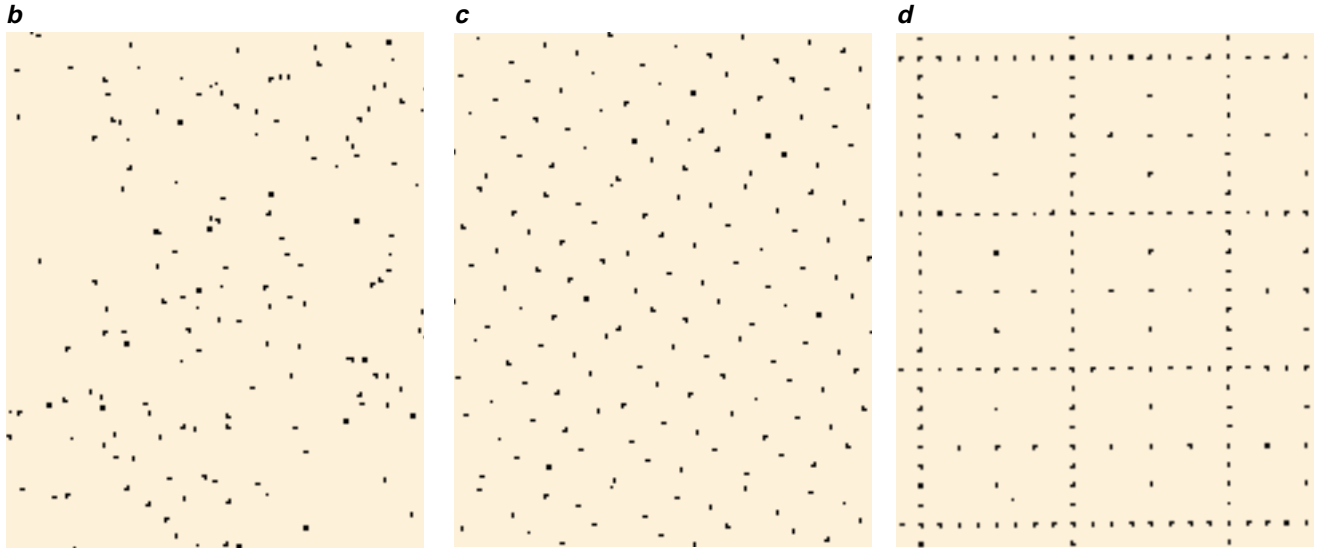
In the 1940s physicists working on the Manhattan Project at Los Alamos National Laboratory realized that some of the problems they were trying to solve, such as the movement of neutrons through materials, lay beyond the reach of deterministic calculations. They turned to the Monte Carlo method of Nicholas C. Metropolis and Stanislaw M. Ulam. The strength of the method is that its error does not depend on the number of variables in the problem. Hence, if applicable, it breaks the curse of dimension. The classical Monte Carlo method for multivariate integration requires at most of order $1/\epsilon^2$ evaluations at random points, where ϵ is the error bound. An alternative statement is that if the integrand is evaluated at n random points, then the expected error of randomization is at most of order $1/\sqrt{n}$. Since its formulation, the Monte Carlo method and its variations have proved to be useful to calculate a



Stanislaw M. Ulam, 1909–84

variety of phenomena, from the size of cosmic showers to the percolation of a liquid through a solid.

For multivariate integration, the classical Monte Carlo method is optimal only if the smoothness parameter, r , of integrands is zero. In 1959 the Russian mathematician N. S. Bakhvalov began pioneering research on the computational complexity of multivariate integration in the randomized setting and devised an alternative to the Monte Carlo method. Later, in 1988, Erich Novak of the University of Erlangen-Nürnberg extended the work of Bakhvalov to establish that the computational complexity in the randomized setting is of order $(1/\epsilon)^s$, with $s = 2/(1 + 2r/d)$. Note that $0 < s \leq 2$. If the smoothness parameter equals zero, then $s = 2$, and the classical Monte Carlo method is optimal. On the other hand, if r is positive, then the classical Monte Carlo method is no longer optimal, and Bakhvalov's method can be used instead.



parameter is zero, and the computational complexity becomes infinite. Hence, for many problems with a large number of variables, guaranteeing that an approximation has a desired error becomes an unsolvable or intractable task.

Mathematically, the computational complexity results we have described apply to the so-called worst-case deterministic setting. The “worst case” phrasing comes from the fact that the approximation provides a guarantee that the error always falls within ϵ . In other words, for multivariate integration, an approximation within the error threshold is guaranteed for every integrand that has a given smoothness. The word “deterministic” arises from the fact that the integrand is evaluated at deterministic (in contrast to random) points.

In this worst-case deterministic setting, many multivariate problems are unsolvable or intractable. Because these results are intrinsic to the problem, one cannot get around them by inventing other methods.

One possible way to break unsolvability and intractability is through randomization. To illustrate how randomization works, we will again use multivariate integration. Instead of picking points deterministically or even optimally, we allow (in an informal sense) a coin toss to make the decisions for us. A loose analogy might be sampling polls. Rather than ask every registered voter, a pollster conducts a small, random sampling to determine the likely winner.

Theorems indicate that with a random selection of points, the computational complexity is at most on the order of the reciprocal of the square of the error threshold ($1/\epsilon^2$). Thus, the problem is always tractable, even if the

smoothness parameter is equal to zero.

The workhorse of the randomized approach has been the Monte Carlo method. Nicholas C. Metropolis and Stanislaw M. Ulam suggested the idea in the 1940s. In the classical Monte Carlo method the integrand is evaluated at uniformly distributed random points. The arithmetic mean of these function values then serves as the approximation of the integral.

Amazingly enough, for multivariate integration problems, randomization of this kind makes the computational complexity independent of dimension. Problems that are unsolvable or intractable if computed from the best possible deterministic points become tractable if approached randomly. (If r is positive, however, then the classical Monte Carlo method is not the optimal one; see box on the opposite page.)

One does not get so much for nothing. The price that must be paid for breaking the unsolvability or intractability is that the ironclad guarantee that the error is at most ϵ is lost. Instead one is left only with a weaker guarantee that the error is probably no more than ϵ —much as a pre-election poll is usually correct but might, on occasion, predict a wrong winner. In other words, a worst-case guarantee is impossible; one must be content with a weaker assurance.

Randomization makes multivariate integration and many other important problems computationally feasible. It is not, however, a cure-all. Randomization fails completely for some kinds of problems. For instance, in 1987 Greg W. Wasilkowski of the University of Kentucky showed that randomization does not break intractability for surface re-

Average-Case Complexity

In the text, we mention that the average-case complexity of multivariate integration is on the order of the reciprocal of the error threshold ($1/\epsilon$) and that for surface reconstruction, it is the square of that reciprocal ($1/\epsilon^2$). For simplicity, we ignored some multiplicative factors that depend on d and ϵ . Here we provide more rigorous statements.

The average computational complexity, $\text{comp}^{\text{avg}}(\epsilon, d; \text{INT})$, of multivariate integration is bounded by

$$\frac{g_1(d)}{\epsilon} \left(\log \frac{1}{\epsilon} \right)^{(d-1)/2} \leq \text{comp}^{\text{avg}}(\epsilon, d; \text{INT}) \leq \frac{g_2(d)}{\epsilon} \left(\log \frac{1}{\epsilon} \right)^{(d-1)/2}$$

The average computational complexity, $\text{comp}^{\text{avg}}(\epsilon, d; \text{SUR})$, of surface reconstruction is bounded by

$$\frac{g_3(d)}{\epsilon^2} \left(\log \frac{1}{\epsilon} \right)^{2(d-1)} \leq \text{comp}^{\text{avg}}(\epsilon, d; \text{SUR}) \leq \frac{g_4(d)}{\epsilon^2} \left(\log \frac{1}{\epsilon} \right)^{2(d-1)}$$

Good estimates of $g_1(d)$, $g_2(d)$, $g_3(d)$ and $g_4(d)$ are currently not known.

construction. Is there an approach that does and that works over a broad class of mathematics problems?

There is indeed. It is the average-case setting, in which we seek to break unsolvability and intractability by replacing a worst-case guarantee with a weaker one: that the expected error is at most ϵ . The average-case setting imposes restrictions on the kind of mathematical inputs. These restrictions are chosen to represent what would happen most of the time. Technically, the constraints are described by probability distributions; the distributions describe the likelihood that certain inputs occur. The most commonly used distributions are Gaussian measures and, in particular, Wiener measures.

Although it was known since the 1960s that multivariate integration is tractable on the average, the proof was nonconstructive. That is, it did not specify the optimal points to evaluate the integrand, the optimal combinatorial algorithm and the average computational complexity. Attempts to apply ideas from other areas of computation to determine these unknowns did not work.

For example, evaluating the integrand at regularly spaced points, such as those on a grid, are often used in computation. But theorems have shown them to be poor choices for the average-case setting. A proof was given in 1975 by N. Donald Ylvisaker of the University of California at Los Angeles. It was later generalized in 1990 by Wasilkowski and Anargyros Papageorgiou, then studying for his Ph.D. at Columbia University.

The solution came in 1991, when Woźniakowski found the construction. As sometimes happens in science, a result from number theory, a branch of mathematics far removed from average-case complexity theory, was crucial. Part of the key came from work on number theory by Klaus F. Roth of Imperial College, London, a 1958 Fields Medalist. Another part was provided by recent work by Wasilkowski.

Let us describe the result more precisely. First, put the smoothness parameter at zero—that is, tackle a problem that is unsolvable in the worst-case deterministic setting. Next, assume that integrands are distributed according to a Wiener measure. If we ignore certain

multiplicative factors for simplicity's sake, the average computational complexity has been proved to be inversely proportional to the error threshold (on the order of $1/\epsilon$) [see box on page 105]. For small errors, the result is a major improvement over the classical Monte Carlo method, in which the cost is inversely proportional to the square of the error threshold ($1/\epsilon^2$).

The average case offers a different kind of assurance from that provided by the randomized (Monte Carlo) setting. The error in the average-case setting depends on the distribution of the integrands, whereas the error in the randomized setting depends on a distribution of the sample points. In our books-on-a-shelf analogy, the distribution in the average-case setting might rule out the inclusion of many oversize books, whereas the distribution in the randomized setting determines which books are to be sampled.

In the average-case setting the optimal evaluation points must be deterministically chosen. The best points are Hammersley points or hyperbolic-cross points [see illustration on pages 104 and 105]. These deterministic points offer a better sampling than randomly selected or regularly spaced (or grid) points. They make what would be impossible to solve tractable on average.

Is surface reconstruction also tractable on the average? This query is particularly important because, as already mentioned, randomization does not help. Under the same assumptions we used for integration, we find that the average computational complexity is on the order of $1/\epsilon^2$. Hence, surface reconstruction becomes tractable on average. As was the case for integration, hyperbolic-cross points are optimal.

We are now testing whether the average case is a practical alternative. A Ph.D. student at Columbia, Spassimir H. Paskov, is developing software to compare the deterministic techniques with Monte Carlo methods for integration. Preliminary results obtained by testing certain finance problems suggest the superiority of the deterministic methods in practice.

In our simplified description, we ignored a multiplicative factor that affects the computational complexity. This factor depends on the number of variables in the problem. When the number of variables is large, that factor can become huge. Good theoretical estimates of the factor are not known, and obtaining them is believed to be very hard.

Woźniakowski uncovered a solution: get rid of that factor. Specifically, we say a problem is strongly tractable if the number of function evaluations needed

Discrete Computational Complexity

This article discusses intractability and breaking of intractability for multivariate integration and surface reconstruction. These are two examples of continuous problems. But what is known about the computational complexity of discrete, rather than continuous, problems? The famous traveling salesman problem is an example of a discrete problem, in which the goal is to visit various cities in the shortest distance possible.



A discrete problem is intractable if its computational complexity increases exponentially with the number of its inputs. The intractability of many discrete problems in the worst-case deterministic setting has been conjectured but not yet proved. What is known is that hundreds of discrete problems all have essentially the same computational complexity. That means they are all tractable or all intractable, and the common belief among experts is that they are all intractable. For technical reasons, these problems are said to be NP-complete. One of the great open questions in discrete computational complexity theory is whether the NP-complete problems are indeed intractable [see "Turing Machines," by John E. Hopcroft; SCIENTIFIC AMERICAN, May 1984].

for the solution is completely independent of the number of variables. Instead it would depend only on a power of $1/\epsilon$. The possibility seems too much to hope for, but it was proved last year that multivariate integration and surface reconstruction are both strongly tractable on the average.

A final obstacle must be overcome before these new results can be used. We know there must exist evaluation points and a combinatorial algorithm that make integration and surface reconstruction strongly tractable on the average. Unfortunately, the proof of this result does not tell us what the points and algorithms are, thus leaving a beautiful challenge for the future.

Work on information-based complexity has led one of us (Traub) to speculate that it might be possible to prove formally that certain scientific questions are unanswerable. The proposed attack is to prove that the computing resources (time, memory, energy) do not exist in the universe to answer such questions.

One important achievement of mathematics over the past 60 years is the idea that mathematical problems may be undecidable, noncomputable or intractable. Kurt Gödel proved the first of these results. He established that in a sufficiently rich mathematical system, such as arithmetic, there are theorems that can never be proved.

We believe it is time to up the ante and try to prove there are unanswerable scientific questions. In other words, we would like to establish a physical Gödel's theorem. The process offers a markedly different challenge from proving results about mathematical problems, because a scientific question does not come equipped with a mathematical formulation. Such questions include when the universe will stop expanding and what the average global temperature will be in the year 2001.

Why do intractability results suggest that some scientific questions might be unanswerable? Recall the results. In the worst-case deterministic setting, the computational complexity of many continuous problems grows exponentially with dimension. Also, the computational complexity of many discrete problems is conjectured to grow exponentially with the number of inputs [see box on opposite page]. Furthermore, although some problems are tractable in the randomized or average-case settings, it has been proved that others remain intractable. Such problems may lurk in certain supercomputing tasks or questions regarding the foundations of physics. After all, they involve a large



REENTRY OF SPACE SHUTTLE provides an example of a computationally complex task: modeling of the airflow around the craft. This job is difficult even though only seven variables govern the dynamics. Added dimensions may yield problems that can never be solved and thus limit what is scientifically knowable.

number of variables or particles. Even worse, many physics problems require solutions to a kind of integral called a path integral, which has an infinite number of dimensions. Solutions of path integrals invite high-dimensional approximations. Thus, the intractability results and conjectures are certainly daunting because they suggest that many tasks with a large number of variables or objects might be impossible to solve.

We emphasize the possibility of other impediments to answering scientific questions. One is chaos, the extreme sensitivity to initial conditions. Because the precise initial conditions are either not known or cannot be exactly entered into a digital computer, certain questions about the behavior of a chaotic system cannot be answered. To focus on the issue at hand, we limit ourselves to intractability.

As we have already observed, a scientific question does not come equipped with a mathematical formulation. Each of a number of models might capture the essence of a scientific question. Because intractability results refer to a particular mathematical formulation, it might happen that although a particular mathematical formulation is intractable, another formulation may be found that is indeed tractable. This prospect indicates a possible way to prove the existence of unanswerable scientific questions. We can attempt to show that there exist scientific questions such that every mathematical formulation that captures the essence of

the question is intractable. We would therefore have science's version of Gödel's theorem.

Humans are intrigued not only by the unknown but also by the unknowable. Here we have suggested one possible attack to establish what may be forever unknowable in science. The curse of dimension, broken now for many kinds of problems, may yet cast its spell.

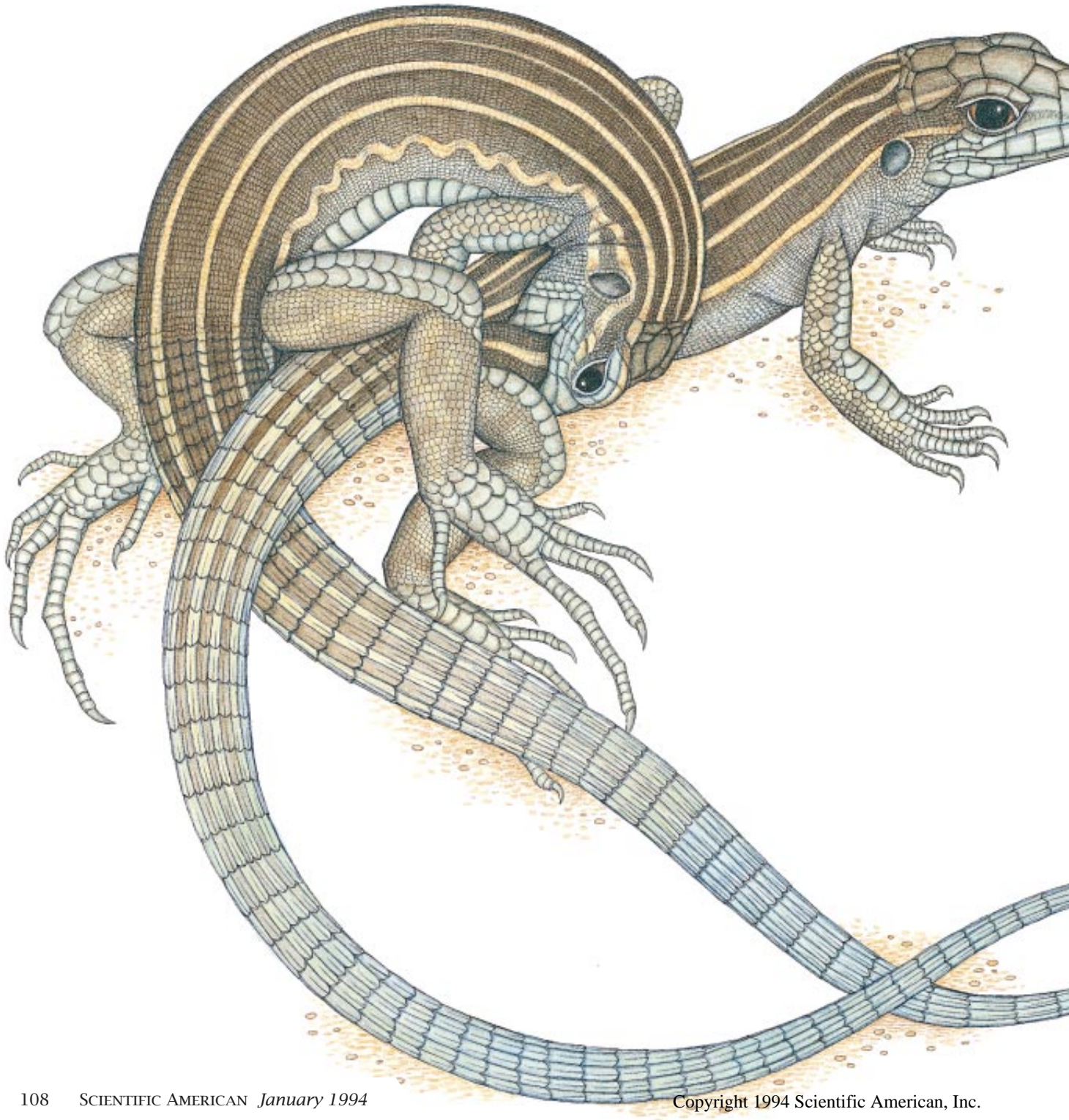
FURTHER READING

- INFORMATION-BASED COMPLEXITY. E. W. Packel and J. F. Traub in *Nature*, Vol. 328, No. 6125, pages 29-33; July 2, 1987.
- INFORMATION-BASED COMPLEXITY. J. F. Traub, G. W. Wasilkowski and H. Woźniakowski. Academic Press, 1988.
- AVERAGE CASE COMPLEXITY OF MULTIVARIATE INTEGRATION. H. Woźniakowski in *Bulletin of the American Mathematical Society*, Vol. 24, No. 1, pages 185-194; January 1991.
- THE COMPUTATIONAL COMPLEXITY OF DIFFERENTIAL AND INTEGRAL EQUATIONS: AN INFORMATION-BASED APPROACH. Arthur G. Werschulz. Oxford University Press, 1991.
- THEORY AND APPLICATIONS OF INFORMATION-BASED COMPLEXITY. J. F. Traub and H. Woźniakowski in *1990 Lectures in Complex Systems, Santa Fe Institute*. Edited by Lynn Nadel and Daniel L. Stein. Addison-Wesley, 1991.
- WHAT IS SCIENTIFICALLY KNOWABLE? J. F. Traub in *Carnegie Mellon University Computer Science: A 25th Anniversary Commemorative*. Edited by Richard F. Rashid. Addison-Wesley, 1991.

Animal Sexuality

Animals have evolved a range of mechanisms to determine whether an individual takes on masculine or feminine traits. Cross-species comparisons offer some surprising insights into the nature of sexuality

by David Crews



One of the most fundamental characteristics of life is sexuality, the division into male and female. Sexual considerations influence the appearance, form, behavior and chemical makeup of nearly all multicellular organisms. Amazingly enough, scientists cannot conclusively say why sex exists. In recent years, however, animal studies have provided a great deal of information about the multifaceted components of sexuality. These studies reveal that many familiar aspects are less universal than once supposed. The work provides a new framework for understanding the relationship between males and females and a glimpse at how sex evolved.

Among vertebrate animals, sexuality is expressed in a number of ways. Males and females exhibit a wide variety of chemical, anatomic and behavioral disparities. The most obvious of the behavioral divergences lies in an animal's copulatory activity. In general, individuals having testes attempt insemination (male-typical behavior), whereas individuals having ovaries are receptive to being inseminated (female-typical behavior). Males and females often differ in other, less overt ways, such as level of activity, regulation of body weight, level of aggression and learning patterns. Some gender-specific actions are associated with, but not necessarily caused by, systematic dissimilarities in certain parts of the brain.

Over the past four decades, biologists have pieced together a master outline of the nature of sexuality, known as the organizational concept. Although it is not totally inclusive, the organizational concept broadly accounts for the structure of sexuality in humans and other mammals. A number of my colleagues and I are currently investigating how to apply the outline more generally to all vertebrate animals.

According to the organizational concept, an animal's sex—specifically, the nature of its gonads—is determined at the time of conception by the chromosomal constitution inherited from its parents. The gonads produce sex ste-

roid hormones that circulate during the early stages of embryonic development; these hormones sculpt the individual's masculine or feminine features. Male sexual traits are instigated primarily by androgens, a class of hormones (including testosterone) produced in the testes. Individuals that lack testes develop ovaries, which generate mostly female hormones called estrogens and progesterins. In this scenario, the female is the neutral, or default, sex, whereas the male is the organized sex.

A key element of the organizational concept is the central role of sex steroid hormones. Modern understanding of the influence of such hormones on sexual differentiation began with the work of Frank R. Lillie of the University of Chicago. Early in this century, Lillie observed that when cows gave birth to twins of opposite sexes, the female twin was sterile and had masculine traits. Lillie, who was an embryologist by training, suggested that androgenic hormones secreted by the male twin in the womb imbued the female twin with some male traits. Scientists have since thoroughly corroborated Lillie's conjecture that the gonads in embryos secrete the hormones that cause males to differ from females.

Among mammals, an embryo starts out having a mass of primordial sexual tissue. Genetic signals determine whether that tissue develops into male or female gonads. Subsequent hormonal triggers that act in the embryo control the sex of the genitalia. The testes of genetic males produce significant concentrations of androgens, which induce the formation of the vas deferens, a penis and a scrotum. In the absence of androgens, the embryo acquires female sexual organs: a uterus, a clitoris and vaginal labia.

Accumulating evidence from animal experiments suggests that many components of adult sexuality—not just the structure of the sexual organs—depend on the hormonal environment during fetal development. Some of the most persuasive support of this notion comes from studies of species that produce litters of many young from each pregnancy. During such pregnancies, the fetuses are arranged like peas in a pod inside the uterus. This grouping results in female and male fetuses residing next to one another in random order.

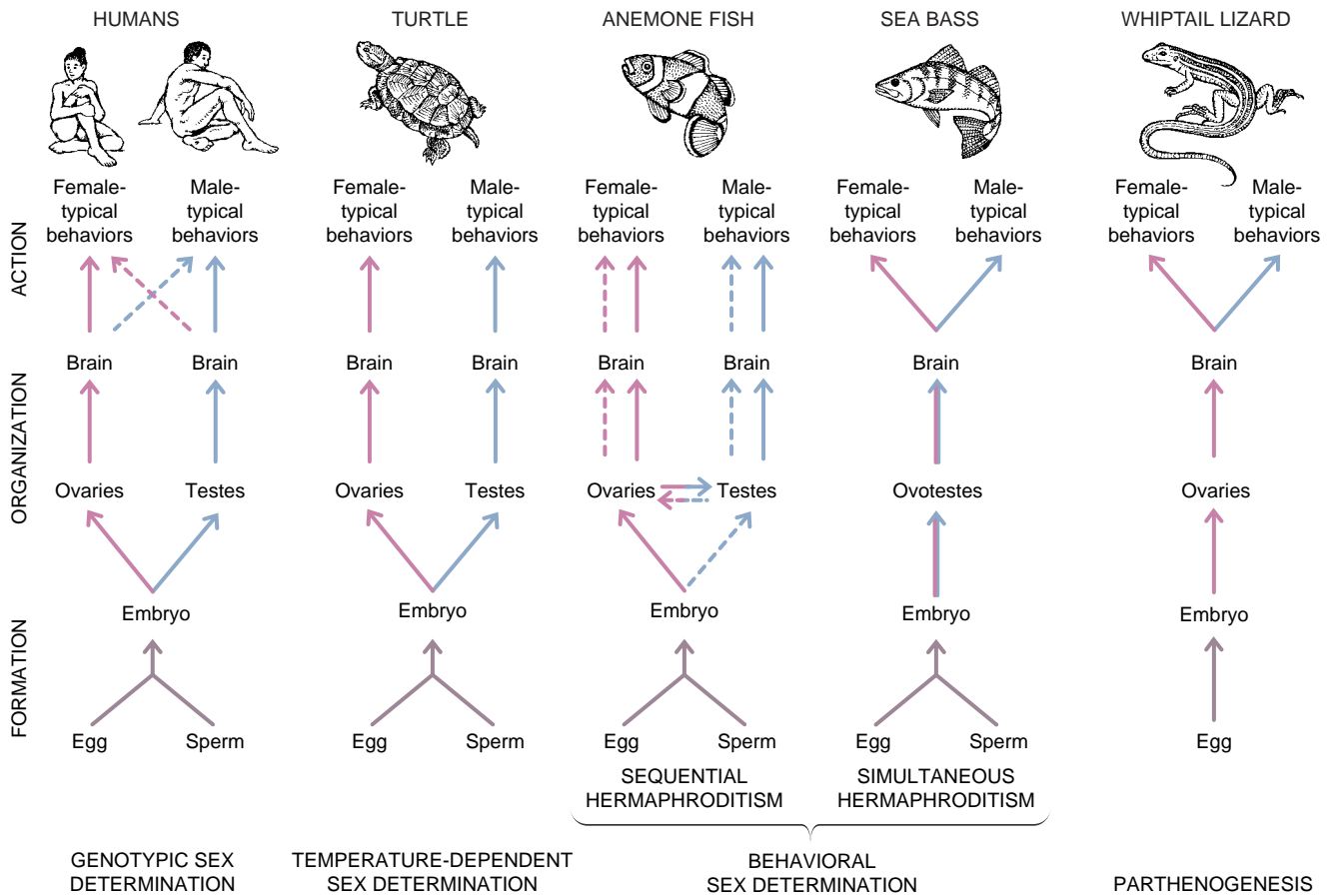
In such an environment, steroid hormones produced by one fetus's gonads could influence the developing neural and secondary and accessory sex structures in an adjacent fetus. Lynwood G. Clemens of Michigan State University discovered that the hormonal surroundings created by neighboring fetuses can profoundly affect adult sexuality in rats. Mertice Clark and Bennett G. Galef of McMaster University have recently observed a similar effect among gerbils.

Frederick S. vom Saal of the University of Missouri has conducted an especially thorough study of sexual development in mice. He found that female mice developing between two male mouse fetuses (known as 2M females) are exposed to higher concentrations of testosterone and lower concentrations of estrogen than are female fetuses that do not develop next to a male (0M females). After birth, 2M females evince a more masculine anatomy than do 0M females. 2M females also take longer to reach puberty and have shorter and fewer reproductive cycles as adults. Finally, compared with 0M females, 2M females are less attractive and less sexually arousing to males and are more aggressive toward other females.

The spotted hyena, a nocturnal African carnivore, offers another prime example of how fetal hormones can direct adult sexuality. Female hyenas exhibit many characteristics usually associated with male mammals. Adult female hyenas within a clan, or social group, are larger and heavier than the males; females dominate nearly all of the adult males in aggressive disputes and in access to food. Female spotted hyenas have normal-looking ovaries and internal genitalia, but their external geni-

DAVID CREWS has spent years exploring the evolutionary roots of sexuality and the role of hormones in controlling sexual differentiation and behavior. He received a Ph.D. in biology from Rutgers University in 1973. In 1982 he moved to the University of Texas at Austin, where he is currently a professor of zoology. Crews's research has focused on sexuality among garter snakes, whiptail lizards and, most recently, red-eared slider turtles and leopard geckos. Crews founded Reproductive Sciences, a biotechnology firm that is using a patented process of hormone-induced sex determination to assist in the breeding of ostriches, emus and other rare birds. In 1992 he also formed Reptile Conservation International, a nonprofit corporation that is using estrogen treatment selectively to build up female populations of endangered turtles in Brazil and Mexico.

WHIPTAIL LIZARDS engage in elaborate mating rituals even though some whiptail species consist of self-reproducing females only. Sexual behavior seems to be a deeply ingrained trait that serves biological functions other than just fertilization; for example, mounting induces asexual whiptails to lay more eggs.



SEXUAL DIFFERENTIATION occurs in all vertebrate species but through several quite different mechanisms. In mammals, chromosomes inherited at the time of fertilization dictate whether an individual develops male or female sexual organs. In many reptiles, incubation temperature of the embryo controls an individual's sex. Hermaphroditic animals

switch from male to female reproductive behavior, usually triggered by the individual's social environment. Simultaneous hermaphrodites can alternate gender repeatedly. Sequential hermaphrodites change once from male to female, or vice versa. Even parthenogenic species display male- and female-typical sexual behavior.

talia have a strongly masculine morphology. They lack external vaginas, and their labia are fused, forming a scrotal sac complete with two bulging pads of fat that simulate testes. The large, erectile clitoris of a female spotted hyena closely resembles a male's penis. Much like many male animals, female spotted hyenas use their clitorises in greeting displays and in dominance interactions.

Stephen E. Glickman and Laurence G. Frank of the University of California at Berkeley recently deduced that this masculinization occurs in the womb as a consequence of the high levels of the chemical androstenedione in the mother's bloodstream. Androstenedione is an inactive compound that can be converted into either estrogen or testosterone. In the placenta of a pregnant hyena, little of the androstenedione turns into estrogen, which leads to high levels of testosterone in the fetus. The abundant testosterone presumably causes the masculine traits of the female hyenas.

Evidently, some mechanism enables

the hormonal environment of an embryo to influence that animal's adult sexual behavior. In 1959 Charles H. Phoenix, Robert W. Goy, Arnold A. Gerall and William C. Young, while working at the University of Kansas, proposed that steroid hormones secreted in mammalian embryos help to organize the sexuality of the brain. Subsequent research has shown that in vertebrates steroid hormones act directly on specific neurons that are linked together in circuits. These neural circuits seem to provide the impetus for behavioral differences between males and females.

Several recent discoveries greatly clarify the link between hormones, brain structure and sexual behavior. For instance, Pauline I. Yahr and her colleagues at the University of California at Irvine identified a nucleus in the gerbil brain that is present only in males. This nucleus lies embedded in an area that helps to control copulatory behavior in male gerbils. Female gerbils injected with androgen early in life develop this "male" nucleus and take on

some male behavioral characteristics.

Certain species of small songbirds also manifest hormone-influenced brain structures that seem to correspond to gender roles in courtship. Male canaries begin to sing in the spring, when their androgen levels are high. The singing both establishes breeding territories and attracts females. Females respond to the song but do not sing themselves. Fernando Nottebohm of the Rockefeller University and others have determined that the contrasting behavior of male and female canaries and other songbirds is matched by differences in the structures of their brains.

The workers find that singing is mediated by an interconnected series of brain nuclei that control the vocal organs. The song-control regions in the brains of female songbirds normally are much smaller than those in the brains of the males. Steroid hormones in songbird embryos determine which neurons survive and which die. The result is that the size and number of neurons, as well as the quantity of synapses

es, in the song-control nuclei are much greater in males than in females.

Nottebohm has shown that the song-controlling brain nuclei vary in size seasonally, waxing and waning in conjunction with the flow of the reproductive cycle. By castrating male songbirds (to lower their androgen levels) or injecting them with androgen (to raise those levels), he and his colleagues have artificially re-created such seasonal changes in singing. In related work, female canaries given appropriate injections of androgen have been induced to sing [see "From Bird Song to Neurogenesis," by Fernando Nottebohm; SCIENTIFIC AMERICAN, February 1989].

A particularly exciting and controversial discovery of a link between sexual behavior and brain structure concerns homosexuality in humans. Simon LeVay, then at the Salk Institute for Biological Studies in San Diego, has reported that the size of a nucleus in the anterior hypothalamus of homosexual men more closely resembles the comparable structure in women than that in heterosexual men. Dean H. Hamer and his colleagues at the National Institutes of Health claim to have found a region on the X chromosome that may contain a gene or genes for homosexuality. If so, the associated brain structure may be under direct genetic control. It is also possible, however, that the hormonal environment surrounding the fetus may partially or totally control the development of the brain nucleus.

These discoveries illustrate the inadequacy of stereotypical divisions of male or female. As the organizational concept makes clear, sexuality depends on subtle hormonal controls, not just on either-or genetic labeling. This finding applies to all the tissues associated with reproduction, including the circuits in the brain that underlie sexual behavior.

In most vertebrate species, adults usually exhibit mating behaviors characteristic of their own gonadal sex, known as homotypical sexual behaviors. Not infrequently, however, individuals also perform behavior patterns normally associated with the opposite sex, known as heterotypical behaviors. For example, females sometimes engage in mounting, and males sometimes solicit being mounted.

Such heterotypical sexual behaviors are a frequent and important part of the social biology of many species, especially among mammals. Female cows commonly mount other females, a practice that seems to help synchronize the reproductive cycles of the herd. In rhesus monkeys, mounting functions as an indicator of dominance and so main-

tains an orderly social hierarchy. Even though embryonic hormones direct neuronal development, it seems that the brain never completely loses the dual circuitry that permits both homotypical and heterotypical sexual behavior.

So far the organizational concept seems to offer a complete framework by which to understand animal sexuality. There is, however, a danger in making sweeping statements about its nature on the basis of observations of a very small number of species, all of them warm-blooded vertebrates, such as birds and mammals. To evaluate the resulting conclusions about sexuality, one must look at a far more comprehensive range of vertebrate species. Much of my own research has concentrated on determining how well the organizational concept applies to cold-blooded reptiles and fish.

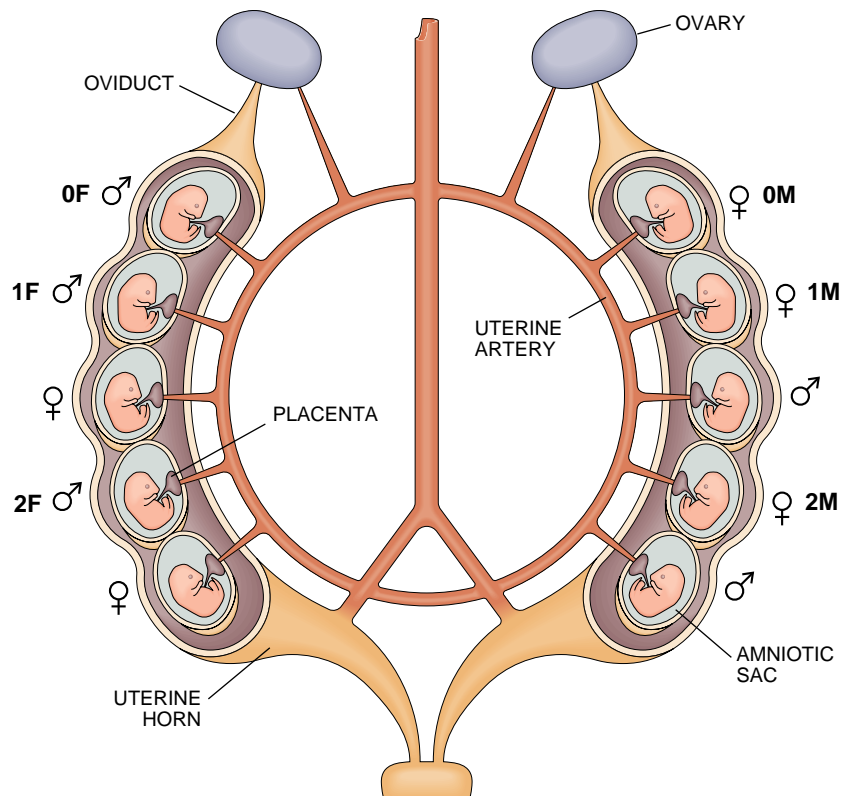
Such investigations are crucial for elaborating a more complete picture of animal sexuality. The kinds of sexual maturation and behavior found in any particular mammal or bird may reflect the unique adaptations of that species. Sexual traits that are shared by many different kinds of vertebrates, in con-

trast, presumably date from a more ancient evolutionary stage. Likewise, sexual behaviors (such as mounting) that appear in both males and females may predate more familiar, sex-specific mating activities. Only by knowing the evolutionary roots of sexuality will scientists be able to learn the rules that govern this omnipresent life process.

When viewed through a broadened perspective that embraces the full diversity of vertebrate animals, the organizational concept clearly fails. I see only one truly essential component of animal sexuality: the organizing effects of sex steroid hormones on the tissues that mediate reproduction. The mechanism that directs those effects varies considerably, however.

In the organizational concept, the sex chromosomes exert ultimate control over whether an animal develops into a male or a female. Yet many fish and reptiles lack sex chromosomes. These species depend on nongenetic triggers to guide sexual differentiation.

Among such species, an individual's gender usually depends on the environment it experiences. In some cases, the determining factor is the temperature at which the embryo develops (tem-



HORMONAL ENVIRONMENT in the uterus affects adult sexuality in mice, gerbils and rats. Female embryos surrounded by males on both sides (2M females) are exposed to higher levels of testosterone than those that do not develop next to a male (0M females). Mature 2M females have a masculinized anatomy; they are also more aggressive and less attractive to males than are 0M females. The opposite, feminizing effect is seen in males surrounded by females (2F males).

perature-dependent sex determination). In other instances, the adult's social surroundings control its sex (behavior-dependent sex determination). Certain animal species even dispense with sexual differentiation and reproduce asexually, a process known as parthenogenesis. These nongenetic methods of sexual differentiation may be evolutionary precursors of the chromosomal control used in mammals.

Temperature-dependent sex determination was identified more than 25 years ago, when Madeline Charnier of the University of Dakar in Senegal reported that the temperature at which rainbow lizard eggs are incubated governs that animal's sex ratio. In the late 1970s James J. Bull and Richard Vogt, then at the University of Wisconsin, conclusively demonstrated that temperature activates some as yet unknown sex-determining mechanism.

Scientists now know that temperature controls gender in many kinds of reptiles, including all crocodylians, many turtles and some lizards. Although all temperature-dependent reptiles lack sex chromosomes, their gender, once set, remains permanent throughout their life. In these species, sex determination occurs in the middle of embryological development, coinciding with the differentiation of the gonads.

Temperature regulation of sexuality takes place in an all-or-nothing fashion. Intermediate temperatures do not produce hermaphrodites; rather they result in a more evenly balanced sex ratio. This pattern indicates that temperature activates a biological switch that determines gonadal sex. I have studied this phenomenon in conjunction with Bull, Judith M. Bergeron of the University of Texas at Austin and Thane Wibbels, now at the University of Alabama at Birmingham. We found that temperature acts by modifying the distribution of enzymes and hormone receptors, including estrogen and androgen receptors, in the growing embryo.

In the leopard gecko, low and high incubation temperatures produce females, whereas intermediate temperatures yield males (different patterns prevail in other species). Working with Bull and William H. N. Gutzke, now at Memphis State University, I administered estrogen to gecko eggs early in development. The estrogen overrode the male-determining temperatures so that all of the young had ovaries.

At temperatures close to those that produce females, lower dosages of estrogen suffice to induce the formation of ovaries. Bergeron, Wibbels and I recently recognized that chemicals that



inhibit the production of estrogens and androgens can prevent an embryo from developing the usual, temperature-controlled male or female gonads. It seems that sex hormones function as the physiological equivalent of incubation temperatures among species that utilize temperature sex determination.

The temperature at which leopard gecko eggs are incubated has a permanent imprint on adult sexuality. Alan Tousignant, a graduate student of mine, and I found that females from eggs incubated at relatively cool temperatures mature faster than those that develop at warmer, predominantly male-producing temperatures. Deborah Flores, another graduate student, and I determined that female geckos from male-biased incubation temperatures are less attractive to males than are females from female-biased temperatures.

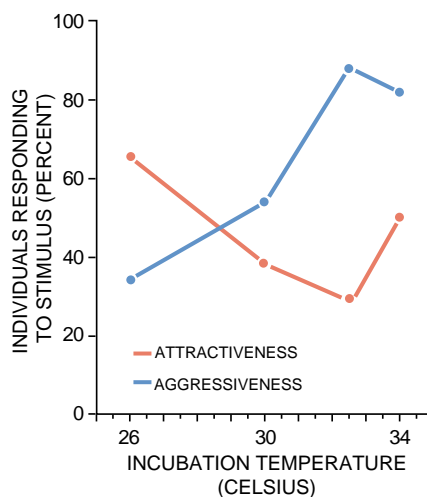
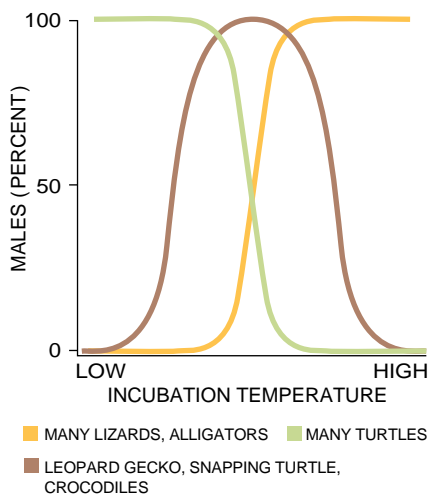
Both male and female leopard geckos are more aggressive if they experienced high temperatures during incubation and are more submissive if they experienced low temperatures. Further, females that incubated at male-biased temperatures develop pubic glands having patent pores similar to those found in males; females that developed at female-biased temperatures have smaller glands and closed pores. In general, the circulating concentrations of androgens in adult females are lower than those in adult males. But the androgen levels in females from male-biased temperatures are higher, and estrogen levels lower, than in those that developed at female-biased temperatures. It seems that adult behavior and sexual chemistry reflect an individual's early, tempera-

ture-modulated hormonal environment.

Temperature-dependent sex determination may be an evolutionary precursor to the genetic control of gender found in mammals. If so, relics of temperature-sensitive behavior might survive in some higher vertebrates. Several workers, including Evelyn Satinoff of the University of Illinois and Christiana L. Williams of Hunter College, report that modifying the temperature drastically influences the behavior of rat pups. This finding hints that even in mammals temperature changes may modulate the organizing effects of steroid hormones. Perhaps the fairly constant body temperatures of warm-blooded animals mask a surviving mechanism whereby temperature can affect the sexual differentiation of the fetus.

Among temperature-dependent species, sex remains fixed once it is set. But species that experience behavior-dependent sex determination, the other main form of nongenetic control of gender, stray even further from the organizational concept and from genetically determined sex. In most cases, these animals are hermaphrodites—that is, individuals that possess both male and female gonads. The social environment controls whether an individual takes on a male or female reproductive role; in other words, sensory stimuli rather than chromosomes direct sexual differentiation. Even so, hermaphroditic species share many chemical and behavioral characteristics with warm-blooded vertebrates.

Some behavior-dependent species of fish are sequential hermaphrodites.



INCUBATION TEMPERATURE of embryos determines the sex ratio in many kinds of reptiles. Depending on the species, the embryos develop into males predominantly at low, intermediate or high temperatures (above, left). Among female leopard geckos, warmer incubation temperatures (up to 32.5 degrees Celsius) engender heightened aggressiveness and reduced attractiveness to males (above, right). The photograph (far left) shows normal courtship among leopard geckos.

These creatures change from one sex to another during their lifetime but express only one gonadal sex at any given time. Orange and white anemone fish are born male and later develop into females. Certain coral reef fish in the Pacific Ocean and in the Caribbean Sea follow the opposite course, starting out female and becoming male. The timing of the sex change depends on a social trigger, such as the disappearance of a dominant male or female.

Other fish species are simultaneously hermaphroditic, meaning they possess a gonad containing both ovarian and testicular tissue. Interestingly, individuals of these species almost never fertilize their own eggs. Instead they continue to mate, perhaps so as to retain the advantages of genetic diversity provided by sexual reproduction. Eric A. Fischer, then at the University of Washington, showed that mating pairs of hermaphroditic butter hamlet alternate between male and female behavioral roles during successive matings. The sex expressed by an individual fish depends on its social surroundings.

How do sequentially hermaphroditic fish accomplish their gender switch? Such species may change from male to female sexual behavior within minutes after witnessing an appropriate change in the number or social structure of the surrounding fish. That rapid transformation must result from signals originating in the brain. Neural connections between the hypothalamus and the gonads exist in all vertebrates. Leo S. Demski of New College of the University of South Florida observed that electrical stimulation in the hypothalamus re-

gion of the brain of the hermaphroditic sea bass can induce the release of eggs or sperm. Perhaps in sequential hermaphrodites these nerves alter the hormonal environment within the gonad; the hormones, in turn, carry the ultimate responsibility for executing changes in sexuality. Less obvious kinds of brain-controlled changes in sexuality may occur in other animals.

Parthenogenesis, or self-cloning, offers yet another alternative to genetically determined reproductive roles. The species that perform this kind of replication consist of females only. One might think that self-cloning species would have no need of any noticeable form of sexual behavior, yet such is not the case. Species of whiptail lizards that reproduce by parthenogenesis exhibit identical mating behavior to related species that engage in conventional sex,

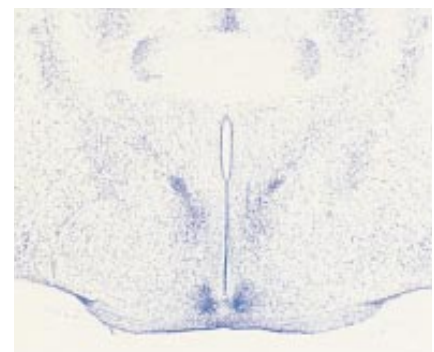
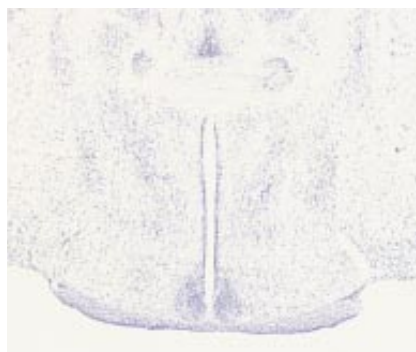
except that each individual alternates between male and female behavior. I have determined that this behavior is controlled by hormones, much as it is in related, sexual species of whiptails.

The persistence of sexual behavior even in an all-female species indicates that such activity is not just a vestigial trait but one that serves an important biological function [see "Courtship in Unisexual Lizards: A Model for Brain Evolution," by David Crews; SCIENTIFIC AMERICAN, December 1987]. Among whiptail lizards, sexual interactions cause the animals to lay many more eggs than they would if they were alone.

My studies of animal sexuality have convinced me that coordinated, complementary, male and female behaviors are crucial for healthy reproduction, even in single-sex species. It is noteworthy that in certain conventional male-and-female species, members of one sex may turn such coordination to their advantage by imitating members of the opposite sex. Such activity may be thought of as another nongenetic form of sexual differentiation.

The bluegill sunfish engages in an intriguing form of such gender bending. Wallace Dominey, while at Cornell University, and Mart R. Gross, now at the University of Toronto, independently discerned that male bluegill sunfish exist in three different forms. Large, colorful males court females and defend their territories. A second kind of male—often known as a "sneaker"—becomes sexually mature at a much younger age and smaller size. These small males live on the periphery of a bigger male's territory and clandestinely mate with females while the dominant male is otherwise occupied.

Sneaker males mature into a third kind of male, one that assumes the behavior and drab coloration of a female sunfish. These female mimics intervene



BRAIN STRUCTURE has been linked to sexual behavior in many species. These images show neural differences in the preoptic region of female (left) and male (right) gerbil brains. That region is involved in the control of masculine copulatory behavior and scent marking, behaviors influenced by androgenic hormones. Females given androgens early in life develop masculine brain structures.



SEXUAL BEHAVIOR does not always cleave neatly into male and female categories. Female spotted hyenas, such as this one with a cub (*left*), have many male behavioral and physi-



cal attributes, including a penislike clitoris used in greeting displays. Pairs of hermaphroditic butter hamlet (*right*) repeatedly trade gender roles during mating.

between a territorial male and the female he is courting. The female mimic, rather than the courting male, usually ends up fertilizing the eggs.

Male red-sided garter snakes enact a similar form of sexual mimicry. At times of peak sexual activity, males congregate around females, forming a so-called mating ball. Robert T. Mason, a former graduate student of mine now at Oregon State University, examined many such balls. He found that in 16 percent of the balls, the snake being courted by the males was in fact a disguised male, which we call a she-male. She-males have testes that produce normal sperm, and they court and mate with females. But in addition to exhibiting male-typical behaviors, she-males produce the same attractiveness pheromone as do adult females. In the mating ball, this second source of the pheromone confuses the more prevalent conventional males, giving the she-male a decided mating advantage.

Numerous studies of lower vertebrates clearly demonstrate that the organizational concept we have outlined here offers an incomplete picture of animal sexuality. I propose that a slightly broader view could encompass all vertebrates. I look beyond the kind of genetically determined sexuality encompassed in the organizational concept toward a more comprehensive, evolutionary view of sexuality. That view builds on the notion that males most certainly evolved only after the evolution of the first self-replicating (and hence female) organisms.

In the organizational concept the female is the default sex and the male the organized sex, imposed on the female by the action of hormones. In my alternative scenario, the female is the

ancestral sex and the male the derived sex. Consider hermaphroditic fishes. Douglas Y. Shapiro of Eastern Michigan University has found that fish species that are born male and become female nevertheless pass through a modified ovarian stage before developing testes. To me, such observations suggest that males may be more like females than females are like males.

Given that every male must contain evolutionary traces of femaleness, biologists might be well served to focus less on the differences between the sexes and more in terms of the similarities. A logical place to concentrate that search would be the sex hormones that are ubiquitous among vertebrates. Some research directed along these lines is in fact paying off. Endocrinologists have found evidence that estrogen and progesterone, both usually associated only with female sexual behavior, may function actively in the sexuality of both genders. In some species, testosterone is converted to estrogen in the brain; in those species, estrogen activates both copulatory behavior in males and sexual receptivity in females. In songbirds, estrogen originates primarily in the brain, implying that its presence transcends gender boundaries and hinting at the existence of brain-controlled sexuality in some higher vertebrates.

Progesterone is generally thought to inhibit sexual activity in males; it even has been used as a form of chemical castration in felony rape convictions. Most researchers therefore have assumed that progesterone has no place in normal male sexuality. Male rats and humans, however, show a pronounced daily rhythm in progesterone secretion; peak progesterone levels occur at the onset of night, when copulatory behavior most often occurs. Diane Witt of the

National Institute of Mental Health, Larry Young, one of my graduate students, and I recently observed that physiological dosages of progesterone can induce castrated male rats to resume mounting. Moreover, injections of RU 486, a hormone that chemically nullifies progesterone, reduces sexual behavior in intact male rats. Like estrogen, progesterone seems to be both a female and a male hormone—an evolutionary relic extending beyond the confines of the organizational concept.

Further investigation of the similarities of males and females may turn up additional instances of “female” aspects of sexuality that might be more correctly viewed as “ancestral.” Such work may yield more clues about the mode of action and evolutionary origins of sex steroid hormones. It may also illuminate connections between the mechanisms by which temperature, brain function and genetics determine gender. In this way, researchers will achieve a deeper and richer understanding of the essential nature of sex.

FURTHER READING

FUNCTIONAL ASSOCIATIONS IN BEHAVIORAL ENDOCRINOLOGY. David Crews in *Masculinity/Femininity: Basic Perspectives*. Edited by J. M. Reinisch, L. A. Rosenblum and S. A. Sanders. Oxford University Press, 1987.

BEHAVIORAL ENDOCRINOLOGY. Jill B. Becker, S. Marc Breedlove and David Crews. MIT Press, 1992.

THE ORGANIZATIONAL CONCEPT AND VERTEBRATES WITHOUT SEX CHROMOSOMES. David Crews in *Brain, Behavior and Evolution*, Vol. 42, Nos. 4-5, pages 202-214; October 1993.

THE DIFFERENCES BETWEEN THE SEXES. Edited by R. V. Short and E. Balaban. Cambridge University Press (in press).

World Linguistic Diversity

The ancestor of each language was taken to its current territory by pioneers, farmers, traders or a conquering elite. Multidisciplinary studies are clarifying their respective roles

by Colin Renfrew

The Greek historian Herodotus reports that Psamitik, a seventh-century pharaoh of Egypt, arranged that two newborn babies should be reared in isolation until their first words together could be heard. Their first recorded utterance was *bekos*. This the pharaoh's scribes discovered to be the word for bread in Phrygian, a language of Anatolia. They concluded that Phrygian was the original language of the earth. Unfortunately, this fanciful experiment seems to have set the standard for later inquiries. By the 19th century speculation on the origin of language had become so vacuous that the Société de Linguistique de Paris banned the subject from its discussions.

Today, at last, advances in archaeology, genetics and linguistics itself are opening a way to a plausible account of the diversity of the world's languages. Many aspects of the problem are still highly controversial, and any attempt at a synthesis can be merely tentative, but the broad features of the process by which languages evolve have begun to be discernible.

History provides a secure foundation for the creation of a reasonable hypothesis. For more than 200 years, linguists have recognized that some languages have such similarities in vocabulary, grammar, the formation of words and the use of sounds that they must stem from a common ancestor. These ancestral alliances they termed language families. The most famous early classifica-

tion of this kind was undertaken in 1786 by Sir William Jones, a British judge at the High Court in Calcutta, who observed relationships between Sanskrit, Greek, Latin, Gothic and Persian. Common words and grammatical features suggested to Jones that the languages had "sprung from some common source." This family is now known as Indo-European.

Subsequent generations have refined and elaborated the analytical methods that Jones employed. At present, the discipline of historical linguistics to which Indo-European research has given rise systematically compares the languages that belong to a family. The comparison enables workers to reconstruct a hypothetical ancestor tongue, called a protolanguage.

This problem of inferring patterns of descent from data observed in the present is also found in evolutionary biology. Biologists have traditionally attempted to reconstruct the genetic relationships among species by studying anatomic and physiological evidence. In recent decades the search has extended to the molecular level. There investigators decipher the line of descent of specific nucleotide sequences in DNA. In each case, the systematic study produces a classification, or taxonomy, entirely based on information currently observable. Such classification is phenetic, or based on overall appearances.

Often the relative similarity of taxonomic units can be diagrammed in the form of a tree. Since Charles Darwin, most workers in historical disciplines, including historical linguistics and paleontology, have tended to equate such a tree with the evolutionary process that led to the current situation. In other words, they have equated the phenetic tree with the phylogenetic tree.

This conflation rests on the strength of several central assumptions. The most important of them is that evolutionary change occurs at a steady, constant rate. As time passes, forms that

had become separated steadily diverge from one another, and innovations in vocabulary arise.

The assumption of a relatively steady rate of change is crucial because differential change obscures the branching pattern. Imagine, for instance, that Danish split off from English and German before those two languages themselves diverged. The true phylogeny would then place English and German on one branch and Danish on another. If German and Danish, however, have altered little, whereas English has changed a lot, a linguist without other points of comparison might mistakenly place German and Danish together, apart from English.

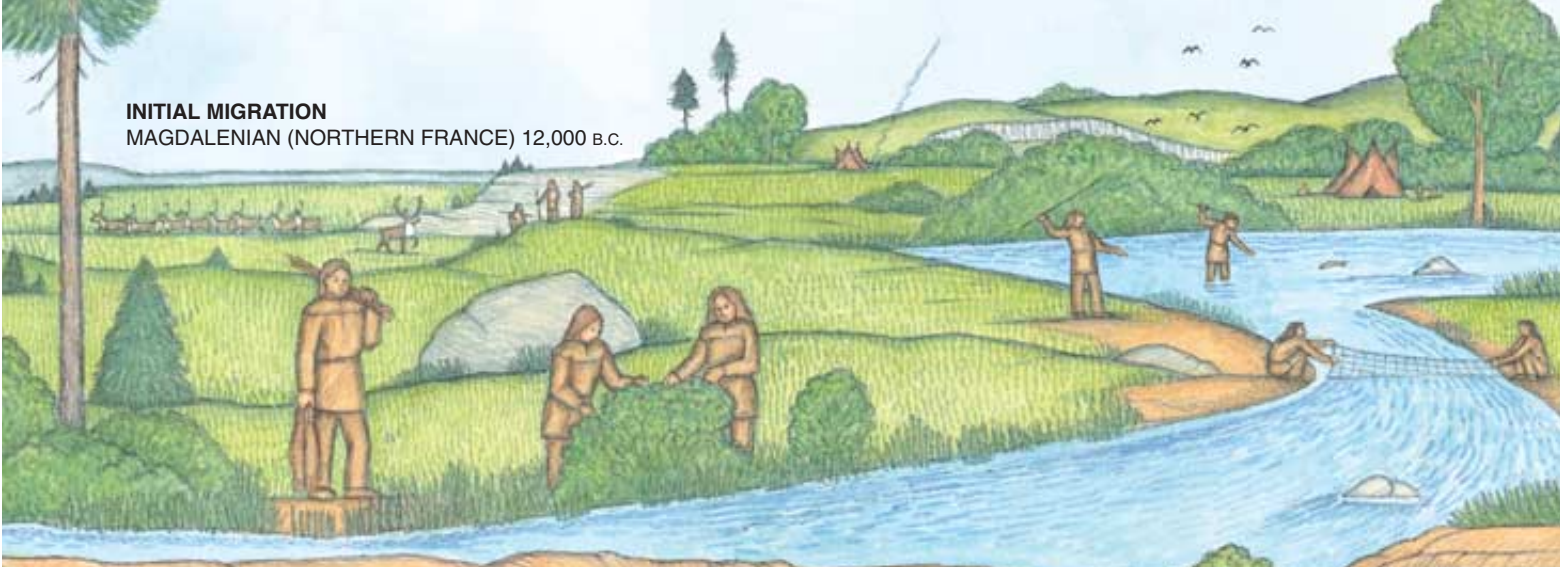
Another assumption is that shared descent, not independent factors that force convergence, accounts for the similarities. In the linguistic context, convergence occurs when contemporaneous languages influence one another through the borrowing of words, phrases and grammatical forms. The almost universal use of the American idiom "O.K." in northern Europe is an example of convergence. Because borrowing rarely affects the most basic elements of a language, workers can usually recognize it. The problem in doing so consists in establishing standards of proof.

Within the discipline of linguistics, enthusiasm for a universal view of the evolution of language is far from unanimous. For many years, it has been possible to recognize two opposing schools of thought among scholars in the field: the "splitters" and the "lumpers." The splitters tend to emphasize the differences that make languages seem unrelated and to split the classification into small, independent units. In their efforts to rule out spurious relationships, split-

LANGUAGES WERE SPREAD by four processes (from top to bottom): initial migrations, demographic expansions of farmers, late incursions into the subarctic, and wide-ranging conquest.

COLIN RENFREW is Disney Professor of Archaeology at the University of Cambridge and Master of Jesus College there; he also directs the McDonald Institute for Archaeological Research. He has conducted excavations in Greece and the British Isles. His previous articles for *Scientific American* were on radiocarbon dating, the obsidian trade, the megalithic monuments of Europe and the origins of Indo-European languages. In 1991 he was made Lord Renfrew of Kaimsthorn.

INITIAL MIGRATION
MAGDALENIAN (NORTHERN FRANCE) 12,000 B.C.



FARMING DISPERSAL
LEVANT 7500 B.C.



LATE CLIMATE-RELATED DISPERSAL
BERING STRAIT 8000 B.C.



ELITE DOMINANCE
GENGHIS KHAN A.D. 1200



ters demand that no group of languages be classed as a family until a series of similarities and affinities has been shown to exist between them. They insist also that these correspondences be used to reconstruct the protolanguage from which the putative family derives. Lumpers, on the other hand, accept criteria that would allow them to lump many languages together into a few families. Although some lumpers also reconstruct protolanguages, others regard this step as superfluous.

Various language families have nonetheless won wide acceptance, among them the Indo-European family; the Afro-Asiatic family (formerly called Hamito-Semitic), which comprises the Semitic languages and most of the languages of North Africa; and the Uralic family, which includes Finnish and Hungarian. The legitimacy of other groupings, however, is far less clear.

In 1963 the American linguist Joseph H. Greenberg of Stanford University took a significant step toward a unified view by classifying the languages of Africa into just four dominant macrofamilies: the Afro-Asiatic, the Khoisan, the Niger-Kordofanian and the Nilo-Saharan. In fact, he did not undertake the historical reconstruction by means of the comparative method, which many linguists would prefer; instead he operated by a system of multilateral analysis. This method simultaneously examines a number of words in many languages rather than comparing words in just a pair of languages.

Despite the reservations of the splitters, Greenberg's classification for Africa has been followed by many scholars. More recently, he has applied the same procedure to the languages of the Americas, identifying three important families or macrofamilies [see "Linguistic Origins of Native Americans," by Joseph H. Greenberg and Merritt Ruhlen; *SCIENTIFIC AMERICAN*, November 1992]. Two of them, the Eskimo-Aleut and the Na-Dene, have found broad support, although his residual category, "Amerind," which incorporates most of the native languages of the Americas into a single macrofamily, has been widely criticized in what has at times been a sharp, even excoriating, debate.

As an archaeologist, I prefer initially to withhold judgment regarding the validity of these macrofamilies, as well as a number of others that the independent linguist Merritt Ruhlen, an undoubted lumper, has advocated. Instead I simply place quotation marks around the controversial ones [see map on opposite page], leaving the question of their nature open while attempting to

solve a more concrete puzzle: How did this distribution come about?

In recent years, suggestions of an answer have come from two archaeological advances, one bearing on the evolution of our species, the other on the evolution of our culture.

The early hominids are now much better understood than they were 20 years ago. No one doubts that it was in Africa, some four or five million years ago, that *Australopithecus* emerged. In Africa, too, some 1.6 million years ago, there developed the ancestor of us all, *Homo erectus*, who dispersed to Asia and Europe and whose fossils and artifacts have been found on both continents. Our own species, *H. sapiens*, certainly split off from *H. erectus* and reached its present form—*H. sapiens sapiens*—more than 100,000 years ago.

Most archaeologists now agree that this process of emergence took place exclusively in Africa. An alternative theory holds that the process of transition from *H. erectus* to *H. sapiens* was not restricted to Africa; rather it took place over a greater area, including Asia and perhaps Europe. But the genetic evidence favors at present the "out of Africa" theory. Following this concept, then, we can envisage the emergence of *H. sapiens sapiens* in Africa about 100,000 years ago and the gradual dispersal of our species through the Old World. By 40,000 years ago modern people had colonized the Levant, southern Asia, Europe, central and eastern Asia, New Guinea and Australia. By perhaps as early as 37,000 years ago—and no later than 16,000 years ago—Asian pioneers had crossed the Bering Strait, beginning the settlement of the New World. We must assume that all these people were speaking a language or languages, even if we may have no clear idea what these languages were like.

The second recent archaeological development of relevance is the emphasis on the mechanisms of cultural change. In particular, archaeologists are no longer willing to explain every alteration in early human culture as the result of some ill-defined migration. They have abandoned the simplistic equation between a language, a culture and a "people." If a migration is to be used as an explanation for a change in decorative art, the appearance of a new religious system or the emergence of a new language, there must be some evidence for the relationship and some understanding of the economic and social processes that gave rise to it.

Four principal processes exist by which a language can come to be spoken in a given territory: initial coloniza-

How Languages Spread

INITIAL MIGRATION

Early humans appear to have spread from Africa to much of the rest of the world beginning about 100,000 years ago. Surviving linguistic traces of this migration include Basque, Caucasian, Khoisan, Australian, "Indo-Pacific" and "Amerind."

FARMING DISPERSAL

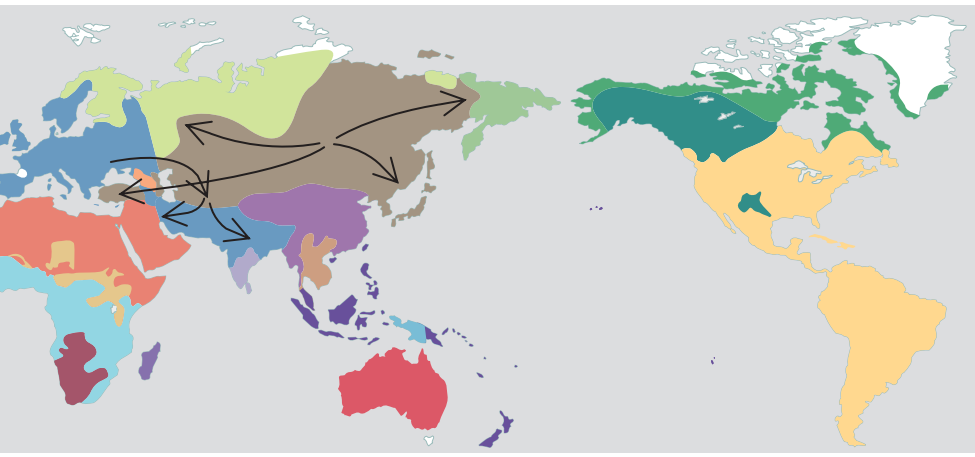
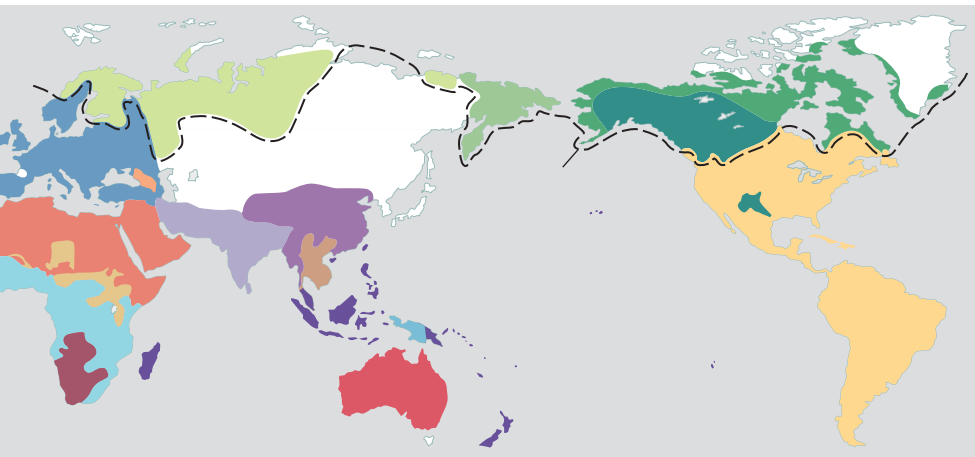
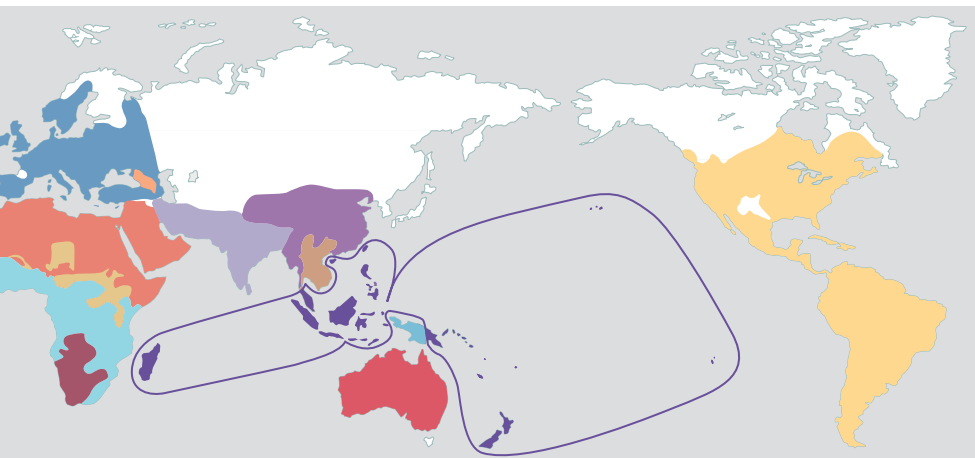
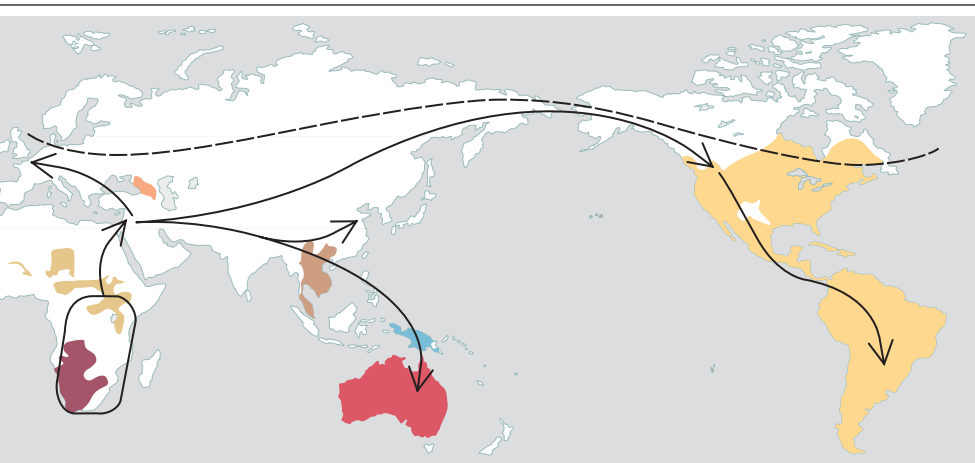
The invention of farming in several places caused populations there to expand. The original farmers' languages therefore spread and differentiated to form such major families as Indo-European, Sino-Tibetan, Austronesian and Afro-Asiatic.

LATE CLIMATE-RELATED DISPERSAL

Global warming several thousand years ago opened regions north of the 54th parallel to pioneers whose languages developed into the families known as Uralic-Yukaghir, Chukchi-Kamchatkan, Eskimo-Aleut and Na-Dene.

ELITE DOMINANCE

The development of complex societies enabled incoming minorities to conquer other populations and to impose their languages on them. The Altaic family spread in this fashion, as did individual members of such previously existing families as Indo-European and Sino-Tibetan.



tion of an unoccupied region; divergence (as discussed earlier); convergence (also discussed earlier); and language replacement, whereby one language is replaced by another incoming language.

If replacement had never occurred, divergence would represent the primary cause of change, and the linguistic map of the world might take the form of a mosaic of small language units. Each language would differ markedly from its neighbors, ranking as a separate family or, more accurately, as a language isolate. This mosaic, in fact, is apparent in the aboriginal languages of northern Australia, where a large number of language families occupy a small area. (Farther south in Australia there is a single, embracing language family, Pama-Nyungan. The explanation for its extraordinarily wide distribution is not clear.) This type of mosaic pattern is found among the horticulturalists of New Guinea. And when one studies the aboriginal language maps of California and of parts of South America (while still, for the present, avoiding Greenberg's "Amerind" classification), one sometimes has a similar impression. So, too, in the Caucasus.

But much of the world map is quite different. Large areas of the globe are occupied by single language families, of the kind that could have arisen only

INITIAL MIGRATION

- KHOISAN
- NILO-SAHARAN
- CAUCASIAN
- AUSTRIC (DAIC AND AUSTRO-ASIATIC)
- "INDO-PACIFIC"
- AUSTRALIAN
- "AMERIND"

FARMING DISPERSAL

- NIGER-KORDOFANIAN
- AFRO-ASIATIC
- INDO-EUROPEAN
- ELAMO-DRAVIDIAN
- SINO-TIBETAN
- AUSTRONESIAN

LATE CLIMATE-RELATED DISPERSAL

- URALIC-YUKAGHIR
- CHUKCHI-KAMCHATKAN
- ESKIMO-ALEUT
- NA-DENE

ELITE DOMINANCE

- ALTAIC

through a process of replacement. I suggest three simple reasons for this pattern.

First, a few families have attained their present extent through the influence of elite dominance. In this model, an incoming minority seizes control of the levers of power and sets itself up as an aristocracy, lending such prestige to its language as to induce the native people to adopt it in preference to their own tongues. Because such minority takeovers imply that the incoming group has some centralized organization, such a hypothesis can apply only in later prehistoric or in historic times, when highly ranked societies had come into existence.

For example, in southern China the Chinese language was adopted only in historical times, through the military expansion of the Chinese empire. The spread of Latin throughout much of Europe also conforms to the principle. So does the diffusion of Indo-European languages through Iran, northern India and Pakistan, which may be attributed to the rise of nomad pastoralism in the second millennium B.C. The Altaic languages became dominant in central Asia in medieval times, when mounted warfare swept that region.

Most large-area language families, however, may be regarded as the product of population dispersals of two different kinds, although these, too, occurred after the end of the last ice age, some 10,000 years ago. The dispersals involved the introduction of farming, on the one hand, and the penetration

of uninhabited areas because of climate changes, on the other.

The recent climate-related dispersals tended to populate empty territories north of the 54th parallel, which had not been habitable during the last cold phase of the Pleistocene. The regions now inhabited by speakers of the Eskimo-Aleut languages were probably first occupied only in the past few thousand years. The Uralic-Yukaghir and Chukchi-Kamchatkan languages would have taken up their present territories earlier than this.

The case of the Na-Dene languages seems more complicated. As Greenberg has suggested, they probably came to North America before the Eskimo-Aleut speakers yet long after the initial colonization of the Americas. Their way of life represented, I believe, an early adaptation to the tundra environment. Later, when climate or ecological factors rendered the area less hospitable to them, they moved south. Some speakers of Proto-Na-Dene penetrated as far as Arizona and New Mexico. Elite dominance, amplified by horseback riding, accounts for the presence of the cultures related to this language group throughout much of the continent.

The most important single factor in the development of the large-area language families seems to have been replacement by means of farming dispersal. According to this theory, a language family begins its career as a single tongue spoken by foragers who live in an ecosystem that con-

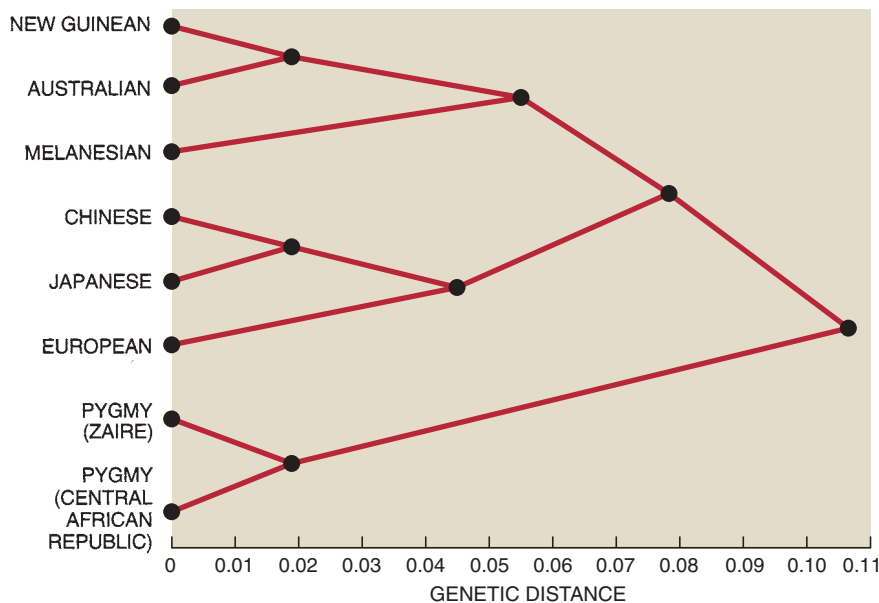
tains plants (and perhaps animals) amenable to domestication. The foragers develop a farming culture that supports them in sedentary habitation, which favors an increased rate of birth, a reduced rate of infant mortality and a greater intensity of food production. The population density increases, assuring the local predominance of the farmers and of their language.

In some instances, the presumption can be made that the domesticated crops and herds, together with the techniques by which they are managed, would prove suitable for transplantation into new ecological niches. In such circumstances, the language or languages of the nuclear area will be transmitted along with the plant and animal domesticates. The languages would move as the farming population expands slowly in a wave of advance, known as demic diffusion. Alternatively, the farmers' language, together with the new agricultural economy, can be adopted by neighboring hunter-gatherer groups through acculturation. The genetic effects of the two mechanisms are significantly different.

It is now accepted by and large that the spread of the Bantu languages of Africa (within the Niger-Kordofanian family) was sustained by demic diffusion. Peter Bellwood of the Australian National University has made the same argument not simply for the Polynesian languages but for the Austronesian languages in general [see "The Austronesian Dispersal and the Origin of Languages," by Peter Bellwood; *SCIENTIFIC AMERICAN*, July 1991].

I have argued this case in some detail for the Indo-European languages of Europe [see "The Origins of Indo-European Languages," by Colin Renfrew; *SCIENTIFIC AMERICAN*, October 1989]. Some authors have argued that in northwestern Europe the process was less one of population movement than of acculturation, yet if that is so, the linguistic effects may have been ultimately the same. Precisely analogous arguments can be advanced for the Afro-Asiatic languages, perhaps for the Elamo-Dravidian languages as well, and for the initial spread of the Altaic languages within Asia. Of course, these languages, especially the Turkic ones, were later carried much farther by the elite dominance of horseback-riding, nomadic pastoralists.

C. F. W. Higham of the University of Otago has recently suggested that comparable arguments would apply to the Austro-Asiatic languages of Southeast Asia (Munda and Mon-Khmer). This group is associated with a Southeast Asian focus on the domestication of rice.



FAMILY TREES derived from gene frequencies constructed for various populations of the world constitute an independent body of evidence with which to compare linguistic, archaeological and anthropological models of prehistory. This tree is based on work done by Joanna L. Mountain and her colleagues at Stanford University.

The Method of Multilateral Comparison

A simple comparison of basic vocabulary reveals such major linguistic groupings as the Germanic, Italic and Slavic branches of Indo-European, Uralic-Yukaghir and Basque.

LINGUISTIC GROUP	LANGUAGE	ONE	TWO	THREE	HEAD	EYE	NOSE	MOUTH
GERMANIC	SWEDISH	en	tvo	tre	hyvud	oga	næsa	mun
	DUTCH	ēn	tve	dri	hōft	ōx	nēs	mont
	ENGLISH	wən	tuw	thrij	həd	aj	nōwz	mawth
	GERMAN	ajns	tsvaj	draj	kopf	augə	nāze	munt
ITALIC	FRENCH	œ/ə	dø	tʁwa	tet	œj	ne	buš
	ITALIAN	uno	due	tre	trsta	okjo	naso	boka
	SPANISH	uno	dos	tres	kabesa	oxo	naso	boka
	RUMANIAN	un	doj	trej	kap	okj	nas	gura
SLAVIC	POLISH	jeden	dwa	trzy	głowa	oko	nos	usta
	RUSSIAN	odin	dva	tri	galava	oko	nos	rot
	BULGARIAN	edin	dva	tri	glava	oko	nos	usta
URALIC-YUKAGHIR	FINNISH	yksi	kaksi	kolme	pää	silmä	nenä	suu
	ESTONIAN	yks	kaks	kolm	pea	silm	nina	suu
BASQUE	BASQUE	bat	bi	hiru	buru	begi	sydur	ahoa

SOURCE: Merrit Ruhlen

The spread of the Sino-Tibetan languages seems initially to have been associated with the domestication of millet and other cereals in the valley of the Yellow River and only later with that of rice.

Naturally, the case for an agricultural expansion of this kind has to be made in detail in each instance. Such inquiries are well within the competence of contemporary archaeology. It is, in fact, generally possible to determine the homeland area of the specific plant or animal domesticates in question and to establish the approximate date of domestication, as well as to document the material record of the dispersal process. The linguistic consequences are of course a matter of inference: prehistoric languages have left no traces in the archaeological record.

The dates for these farming dispersals, increasingly well established by radiocarbon dating, are generally rather earlier than those that linguists have tended to assign for the early phase of the language families in question. Yet the logical basis for the linguistic dating has never been established clearly: no reliable system exists for independently dating protolanguages.

And what of the language families that could not have been spread by people acting on a relatively recent change in climate, a revolution in agriculture or a wave of conquest? Such residual tongues, scattered in bits and pieces throughout the world map, must have arrived in their current ranges

long ago, during the initial dispersal of modern humans. Among these families are the Khoisan and Nilo-Saharan languages of Africa; the northern and southern Caucasian languages; Basque; the Australian languages; the mosaic of perhaps mutually unrelated languages in New Guinea ("Indo-Pacific"); and the pre-Na-Dene languages of the Americas. This last category is so vast that it without doubt embraces several subfamilies whose distributions have been for the most part determined by subsequent processes, including agricultural dispersal.

Molecular genetics can test at least some elements of this overall account of the distribution of languages on the earth. One approach with this method is to compare gene frequencies in various populations and convert these data into a tree, the branches of which represent genetic distance. One can then see to what extent the genetic relationships confirm predictions arising from the above account. Already the out-of-Africa theory for the origin of our own species receives strong confirmation from the family tree based on a sampling of nuclear DNA from a number of living populations [see diagram on preceding page].

Initial dispersals of population into uninhabited territory obviously entail total gene transfer. Agricultural dispersals will involve significant gene flow

only when they proceed through demic diffusion; those that propagate by a process of acculturation will leave fainter genetic traces. Language replacement by elite dominance also involves gene flow on only a very limited scale: usually in such cases, it is the males who travel, so that the effects on mitochondrial DNA (inherited only through the female line) will be minimal.

The most carefully studied case is the coming of farming to Europe, the map of which shows geographic distribution of gene frequencies along a clear gradient from southeast to northwest. Recent statistical work by Robert R. Sokal of the State University of New York at Stony Brook and his colleagues has provided good evidence for associating a significant part of this gradient to the spread of agriculture from Anatolia. Although this correlation supports the view that an expanding population of farmers brought agriculture into new territories, it does not prove that those farmers spoke some of the original Indo-European dialects.

Recently the statistician Guido Barbujani of the University of Padua has conducted a comparable analysis for the other language families whose distribution may be explained by agricultural dispersal from the Levant (such as Afro-Asiatic, Elamo-Dravidian and early Altaic) and has found a similar correspondence. Even more persuasive studies have been carried out in the Pacific, where the spread of the Polynesian lan-

guages correlates impressively with the genetic evidence. In this case, however, the correlation is not surprising, because the Polynesians were occupying uninhabited islands. Therefore, their movement qualifies both as an agricultural dispersal and as an initial dispersal.

Additional supporting evidence from Africa comes from the work of Laurent Excoffier of the University of Geneva and his colleagues. They find a high correspondence between the varieties of gamma globulin in blood samples and the language family of the speakers in question. This is particularly marked for the Afro-Asiatic languages and lends support to the picture outlined here.

The most consistent advocate of the correlation between genetic and linguistic data, however, has been Luigi Luca Cavalli-Sforza of the Stanford University School of Medicine [see "Genes, Peoples and Languages," by Luigi Luca Cavalli-Sforza; *SCIENTIFIC AMERICAN*, November 1991]. In an ambitious exercise he has compared the family tree obtained from molecular genetic data at a world level with a family tree established using only linguistic data. His study indicates a fair degree of overlap.

So far I have adduced no linguistic relationships older than about 10,000 years. Although even this time depth is greater than most linguists would choose to examine, I have justified it not so much by means of new classifications as by proposing unconventionally early dates for well-established language families. Now it is appropriate to go a little further down the path of the lumpers, to note the hypothetical existence of more embracing macrofamilies, such as Amerind and Indo-Pacific. Their origins, presuming in each case a single protolanguage, would probably lie well beyond 20,000 years ago.

Perhaps the best-known macrofamily was worked out by two Russian scholars, the late Vladislav M. Illich-Svitych and Aharon B. Dolgopolsky of the University of Haifa. They have argued that Indo-European, Afro-Asiatic, Dravidian, Altaic and Uralic can be classified together within a single macrofamily they called Nostratic (from the Latin *nostras*, "our countryman"), which is itself derived from a Proto-Nostratic language supposedly spoken in the Middle East some 15,000 years ago. (Greenberg has defined a similar macrofamily, "Eurasianic," which differs by excluding Dra-



DIRECT EVIDENCE of ancient languages begins only some 5,000 years ago, with earliest written records, such as this pictograph inscription from Uruk.

vidian and Afro-Asiatic and including Eskimo-Aleut and Chukchi-Kamchatkan.) Strikingly, these macrofamilies also show a good correlation with the genetic evidence, as marshaled by Cavalli-Sforza, and indeed with some of the archaeological evidence for agricultural dispersals.

Linguistic lumpers have not yet carried the majority of their specialist colleagues with them. Nevertheless, the multilateral analysis method of Greenberg draws on a battery of lexical evidence that is certainly impressive to the nonspecialist. And the Nostratic school does set out to use the comparative method of historical reconstruction, for whose omission Greenberg is so severely criticized by his colleagues. The archaeological and genetic arguments, in fact, harmonize well with some conclusions of the lumpers. The success of the rationales indicates that additional work in this direction will be worthwhile.

Some scholars, notably Ruhlen, have even suggested the existence of much wider underlying affinities between macrofamilies—for instance, between Amerind and Eurasianic. Such a hypothesis holds that some modern word forms demonstrably derive from the single and ultimate protolanguage spoken by our remote African ancestors in their homeland. A claim of this kind is difficult to test and will be rejected by most linguists. Still, linguistic arguments for monogenesis do not contradict the evidence from archaeology, bioanthropology and molecular genetics for an out-of-Africa origin for our species.

These are deep waters. They appear,

however, to convey a glimmer of real historical processes. This assumption is supported by the work of linguists such as Johanna Nichols of the University of California at Berkeley, who analyzes languages according to structural features that may have no genealogical significance. Her interesting recent analysis of structural typology in a large sample of the world's languages has led her to put forth three stages for the origins of the world's languages that could harmonize with the sequence I have suggested here.

She notes the existence of two kinds of language areas. "Spread zones" are large areas occupied by just one or two language families; examples include Europe (with the Indo-European languages) and North Africa (with the Afro-Asiatic languages). "Residual zones" are smaller, although each one harbors a number of long-established language families; examples are provided by the Caucasus and New Guinea. Nichols also sees the spread zones as the result of events that followed the end of the last glaciation; the residual zones are by and large the relics of earlier initial dispersals.

Much more work remains to be done. Nevertheless, a clear convergence is emerging between the archaeological evidence, the genetic evidence and at least some of the linguistic evidence. It would seem, then, that the broad outlines for a major new synthesis are now visible, one that in the coming decade can be expected to clarify not only the diversity of language but also that of genes and cultures.

FURTHER READING

- GENES, PEOPLE AND LANGUAGES. Luigi Luca Cavalli-Sforza in *Scientific American*, Vol. 265, No. 5, pages 104-110; November 1991.
- A GUIDE TO THE WORLD'S LANGUAGES, Vol. 1: CLASSIFICATION, WITH POSTSCRIPT. Merritt Ruhlen. Stanford University Press, 1991.
- ARCHAEOLOGY, GENETICS AND LINGUISTIC DIVERSITY. Colin Renfrew in *Man*, Vol. 27, No. 3, pages 445-478; September 1992.
- LINGUISTIC DIVERSITY IN SPACE AND TIME. Johanna Nichols. University of Chicago Press, 1992.
- WORLD LANGUAGES AND HUMAN DISPERSALS: A MINIMALIST VIEW. Colin Renfrew in *Transition to Modernity: Essays on Power, Wealth and Belief*. Edited by J. A. Hall and I. C. Jarvie. Cambridge University Press, 1992.

The First Data Networks

The optical telegraph is almost forgotten. Two centuries ago it moved messages over hundreds of kilometers in a few minutes

by Gerard J. Holzmann and Björn Pehrson

People usually think of the rise of large-scale communications networks as a 20th-century phenomenon. Many in industrialized countries can remember when telephones first arrived in their neighborhood, and today's young adults will no doubt tell their grandchildren about the days before the Internet. The first nationwide networks, however, were built not in this century but almost 200 years ago. Well before the electromagnetic telegraph was fully developed, many countries in Europe had fully operational nationwide communications systems, each consisting of hundreds of stations.

The first two systems were built in the 1790s by the French clergyman Claude Chappe and the Swedish nobleman Abraham Niclas Edelcrantz. Their passion to develop and build working telegraph systems was not unusual; countless serious and not-so-serious scientists had dabbled at sending messages over long distances since classical times. The remarkable feat was that both men succeeded where everyone else had failed. And they did so under the most challenging circumstances: in the midst of a violent revolution in France and a series of coups d'état in Sweden.



Chappe was 25 years old on July 14, 1789, when the storming of the Bastille raised the curtain on the French Revolution. He had studied for the clergy, but in November he lost his religious benefices and had to return to his place of birth, Brûlon, near Le Mans. In the turmoil of the revolution, his four brothers, Ignace, René, Pierre-François and Abraham, also lost their livelihoods and returned home. Fortunately, Chappe's family was not without means, and Claude began to spend his time on experiments in physics, with the help of his brothers. His investigations soon earned him a membership in the Société Philomatique, a society of physicists in Paris. Between 1789 and 1793 he published five papers on physics and repeatedly touched on the problem of transmitting electrical impulses through wires. Although success had been reported elsewhere, the principles of electricity were not understood well enough to allow anyone to build a practical telegraph.

Perhaps frustrated by his attempts to make a working electrical telegraph, Chappe turned to optical alternatives. On March 2, 1791, he gave a public demonstration of his first system. Each station consisted simply of a modified pendulum clock and a large panel painted black on one side and white on the other. The clockfaces were divided into 10 parts, each used to designate a number. A single hand, or pointer, made one complete rotation of the clockface at least twice a minute.

At the start of a transmission, the sender turned the panel to indicate when the hand of his clock reached the zenith; that allowed the receiver, watching through a telescope, to set the clock on the other end. Subsequent numbers were sent by flipping the panel from white to black each time the pointer of the sender's clock passed over the appropriate position. By looking at the position of the local clock, the receiver could determine what number the sender intended. Messages were encoded with the help of a numbered dictionary of letters, words and phrases. The speed of the telegraph was regulated by the rate at which the clock pointer turned.

The first transmission took place over a distance of roughly 16 kilometers, between the castle in Brûlon and a private house in the neighboring town of Parcé. The local doctor, M. Chenou, chose the first sentence transmitted, at 11 A.M.: "Si vous réussissez vous serez bientôt couvert de gloire" ("If you succeed, you will soon bask in glory"). The message was transmitted in approximately four minutes.

Armed with sworn affidavits from

SEMAPHORE TOWER at Marcy-sur-Anse (opposite page) was built about 1804 as part of a line extending from Paris to Lyons. The tower and its semaphore have recently been reconstructed. The engraving (above) shows a station on the line to Lille, which was built in 1794.

GERARD J. HOLZMANN and BJÖRN PEHRSON studied the history of telecommunications methods independently for many years. They discovered their common interest in 1989 and have collaborated since then. Holzmann works at AT&T Bell Laboratories in Murray Hill, N.J., where he does research in the design and verification of communications protocols, distributed computing and computer graphics. He received his Ph.D. from the University of Technology in Delft, the Netherlands, in 1979. Pehrson is chair of the department of teleinformatics at the Royal Institute of Technology in Stockholm and a member of the board of the Swedish Institute of Computer Science. He received his Ph.D. from Uppsala University in 1975.

the eyewitnesses to his first experiment, Chappe moved to Paris to seek funding for larger tests. Paris was not a quiet town at that time, and on more than one occasion he narrowly escaped with his life when angry revolutionary mobs stormed and destroyed his experimental telegraphs, suspecting them to be part of a royalist plot.

On March 22, 1792, Chappe submitted a formal proposal to the Legislative Assembly. His older brother, Ignace, had meanwhile been elected to that body and could help him gain access. In a brief address to the assembly on March 24, Claude offered an invention that could be used to send "messages, battle orders or anything imaginable" to anywhere in the country, within a matter of minutes.

Nothing happened. Chappe's case was delegated from one committee to another. Finally, on April 1, 1793, Deputy Charles-Gilbert Romme (better known for the introduction of the French republican calendar) intervened on Chappe's behalf. He gave a speech that strongly supported Chappe's work, emphasizing the potential of the invention for military purposes. The new French Republic was at war with most

of its neighbors, so his point was taken. The assembly approved Romme's proposal to fund an experiment. At the same time, a brand-new term was introduced to describe Chappe's device: *télégraphe*, or "far writer." Until then, Chappe had toyed with the term *tachygraphe*, or "fast writer," not exactly an accurate description.

Three telegraph stations were built, with official protection. The first was placed in Paris at Le Peletier Saint-Fargeau Park in Belleville, the second at the heights of Écouen, roughly 16 kilometers to the north, and the third at Saint-Martin-du-Tertre. Chappe had by now abandoned the pendulum design and, after investigating several other options, had settled on a semaphore system. The semaphore consisted of a large horizontal beam, called a regulator. Two smaller beams, referred to as indicators, were mounted at the ends. The array seemed to mimic a person with wide-outstretched arms, holding a signal flag in each hand. The angles of the indicators and the position of the large regulator beam could be varied in increments of 45 degrees, sufficient for the encoding of hundreds of symbols.

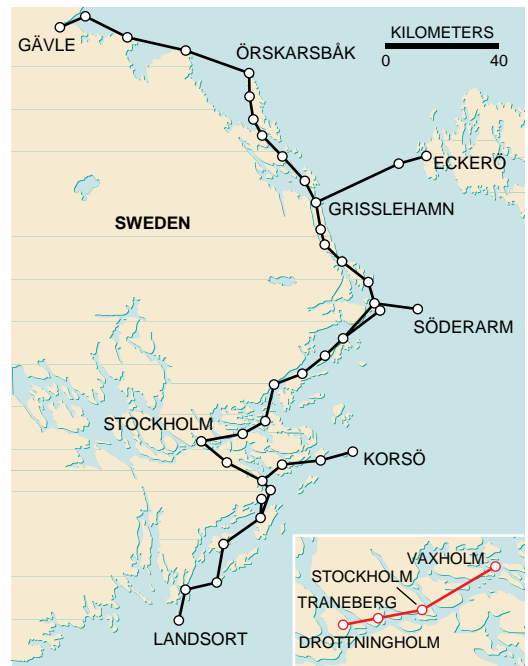
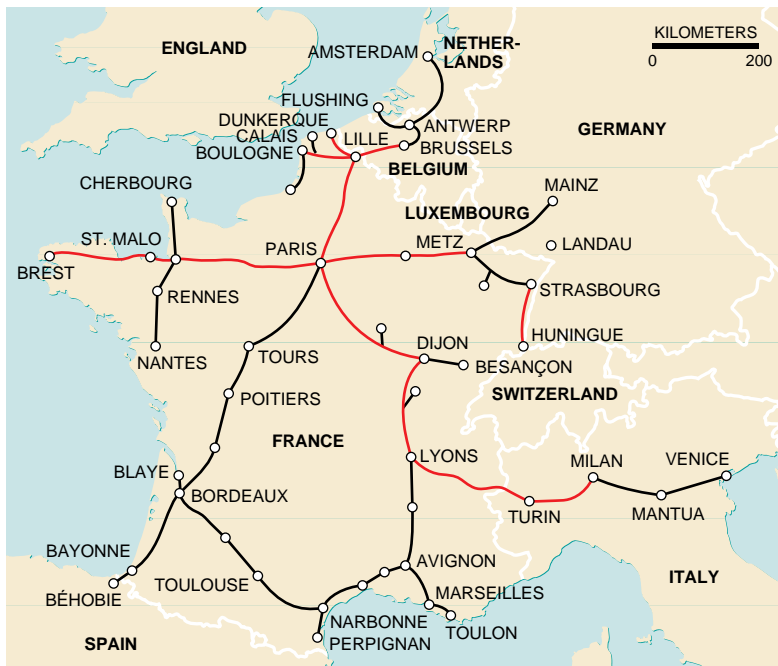
At 4:26 P.M. on July 12, 1793, opera-

tors exchanged the first signals. In 11 minutes, one of the observers, Pierre Daunou, sent the message: "Daunou has arrived here. He announces that the National Convention has just authorized his Committee of General Security to put seals on the papers of the deputies." The answer arrived after nine minutes, presumably the result of more careful thought: "The inhabitants of this beautiful region are worthy of liberty by their respect for the National Convention and its laws."

Some weeks later the National Convention decided to establish a French State Telegraph and to fund the construction of a line of 15 stations connecting Paris to Lille, about 190 kilometers north of Paris, at the frontier with the Austrian Netherlands. Claude Chappe received the title of *Ingénieur Télégraphe*, a salary of 600 francs per month and the permanent use of a government horse. At Chappe's request, his brothers Abraham, Ignace and Pierre-François were appointed as the first administrators of the line to Lille.

Chappe completed the line within a year. On August 15, 1794, the first official message arrived at Paris. This bulletin, announcing the recapture of the





TELEGRAPH NETWORKS grew to cover most of Europe during the decades before 1850. The French network (*left*) was the most extensive (red marks sections built before 1810;

black indicates those built thereafter). The Swedish system (*right*) used shutters instead of a semaphore. (The inset shows the first experimental line, built in 1795.)

city of Le Quesnoy from the Austrians and Prussians, was transmitted within an hour after the battle. The delegates were duly impressed. On August 30 the telegraph again reported happy news: the recapture of Condé. The message read: "Condé être restitué à la République. Reddition avoir eu lieu ce matin à six heures" ("Condé is restored to the Republic. Surrender took place this morning at six o'clock"). More such reports followed as the French advanced north into Holland. Chappe and his telegraph looked better with every victory message that arrived in Paris. On October 3 the legislature decided to extend the telegraph system with a second line from Paris to Landau.

This second line was hampered, as were many projects in those days, by lack of funds; it took four years for the builders to reach Strasbourg. The telegraph office decided to terminate the line there. Instead they extended the line from Paris to Lille by another 64 kilometers to reach Dunkerque. Meanwhile, at the request of the navy, a new line was built to connect Paris to the fleet in Brest (a distance of 210 kilometers). This phase of construction brought the total number of optical telegraph stations in use in 1799 to roughly 150.

When Napoleon Bonaparte seized power later that year, he ordered construction of a 95-kilometer line from Strasbourg to Huningue. In 1803 the telegraphers built lines from Lille to Brussels (96 kilometers) and from Lille

to Boulogne (110 kilometers). The line to Boulogne was constructed with an eye toward a possible invasion of England.

Indeed, two years earlier Napoleon had commissioned the youngest Chappe brother, Abraham, to devise a telegraph that could signal across the English Channel. Abraham designed a large two-arm semaphore and tested it in July 1801 between Belleville and Saint-Martin-du-Tertre, roughly the distance from Boulogne to Dungeness. One semaphore was reportedly constructed in Boulogne, but because the invasion never took place it was quickly abandoned.

In 1804 Napoleon ordered construction of the longest line of the telegraph network, from Paris via Dijon, Lyons and Turin to Milan, a total of 720 kilometers. Within a year the system covered nearly every part of France. Four main branches reached out from Paris to the north, south, east and west, in many cases following the old network of stagecoach routes.

Claude Chappe had reached a pinnacle, although he seems never to have realized it. He suffered from the ever increasing attacks of other inventors, who claimed to have designed telegraphs that either predated or outperformed Chappe's. Toward the end of 1804, he fell ill during a routine inspection tour of some new lines under construction. He suspected that his food had been poisoned and pointed an accusing finger at his adversaries. After several months of convalescence,

Chappe returned to Paris but sank into a depression from which he could not recover. On January 23, 1805, he leapt into the well outside the telegraph administration at the Hotel Villeroi.

The remaining Chappe brothers continued working for the telegraph administration, with the strong support of the government. Napoleon was convinced of the value of the Chappe semaphores and made use of the telegraph in his campaigns. The rapid notification of troop movements may well have helped him outwit his enemies. Indeed, in 1812 Napoleon commissioned Abraham Chappe to develop a mobile semaphore that could be deployed during the invasion of Russia.

While the optical telegraph system was under construction, the method by which its operators transmitted information was evolving as well. The first telegraph code was adapted from one developed in 1791 for the pendulum telegraph. It consisted of a book of 9,999 entries. The first nine, the numerals from one to nine, were encoded in a single signal. The next 89 entries, 10 to 99, were encoded in two signals; those from 100 to 999 required three; and those from 1,000 to 9,999 took four. To speed up transmissions, the most frequently used words and phrases had the lowest numbers (a precursor of modern data-compression techniques).

To change the signaling technique

from synchronized clocks to semaphores, Chappe initially chose a set of simple semaphore signs loosely based on a common shorthand notation. This encoding, however, did not fully exploit the combinations that could be set on the three-armed semaphore and so slowed down transmissions.

In 1795, before construction on the second line to Strasbourg began, Chappe decided to develop a new code optimized for the semaphore. The regulator and the indicator beams on the semaphore could be set at angles that varied in increments of 45 degrees. Each indicator thus had eight different positions, and the regulator had four: in sum 256 possible combinations. Positions where

the indicators either extended the position of the regulator, or were hidden behind it, were almost indistinguishable and were therefore excluded. This change reduced the number of indicator positions to seven each, leaving 196 combinations. To simplify matters further, Chappe restricted the semaphore signs to those for which the regulator beam was placed either horizontally or vertically, reducing the number of combinations to 98. From these, he eliminated another six potentially confusing configurations and adopted the remaining 92 for his code tables.

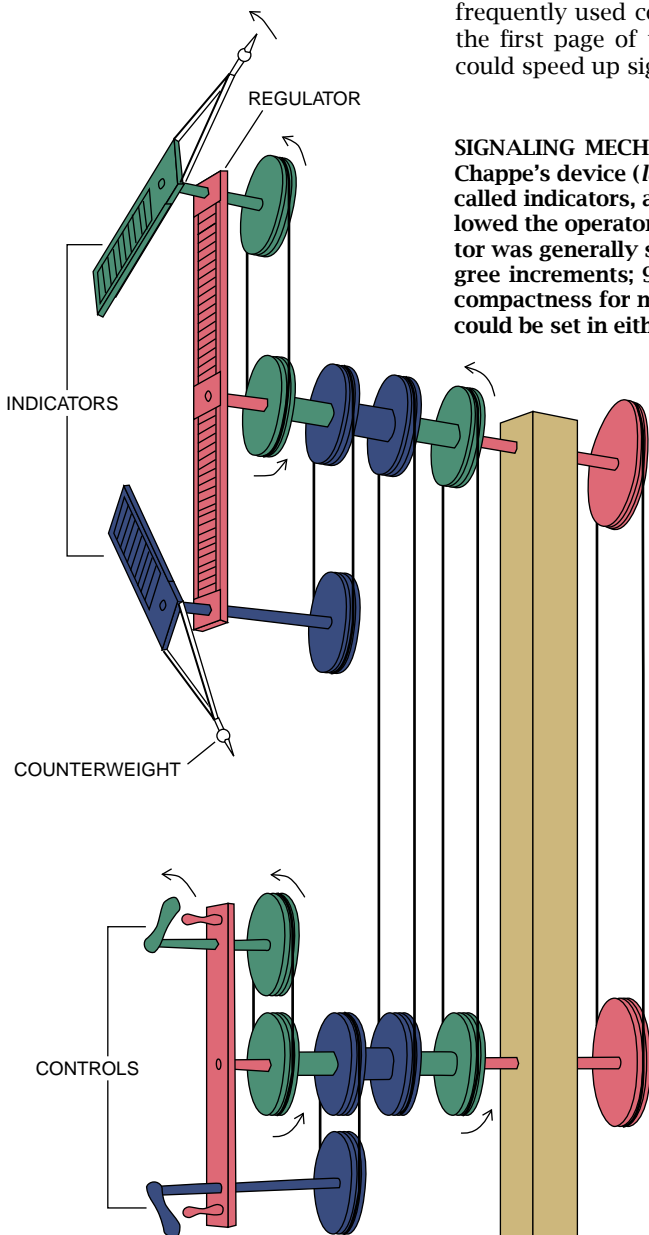
The codebook published in 1795 contained 92 pages of 92 entries, each for a total of 8,464 letters, numerals, words and phrases. Each one was transmitted as a pair of semaphore signals encoding page number and line. The 92 most frequently used codes were placed on the first page of the book. Operators could speed up signaling by indicating

the first page with an abbreviated code called "double-closed," made by folding the indicator wings in from whatever the semaphore position was for the first half of the code pair.

In 1799 Chappe extended the code with two extra books, for a total of 25,392 entries. The additional books contained code pairs for geographical names and common words and phrases. The code also included shift codes to move from one book to another.

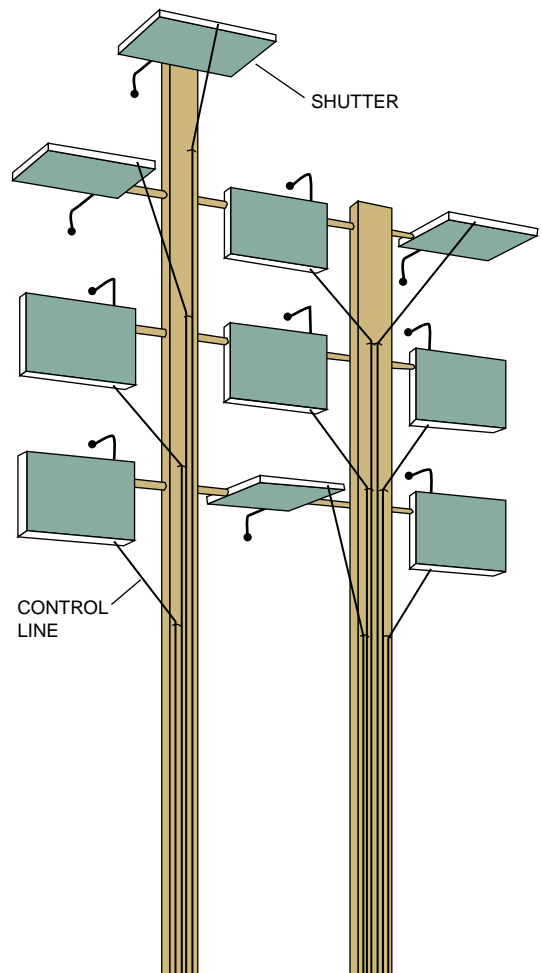
When a conversation began, it took about six seconds to transmit the first signal from one station to another. Transit through the 120 stations between Paris and Toulon, for example, required 12 minutes. In a continuous correspondence between the two cities, passing the signal from station to station required only a second or two, but operators generally held each signal for 20 to 30 seconds to ensure reception. Consequently, only one or two signals

FRENCH OPTICAL TELEGRAPH



SIGNALING MECHANISMS took different shapes in France and Sweden. Claude Chappe's device (*left*) consisted of a long arm, the regulator, with two short arms, called indicators, attached to its ends. A system of belts and concentric shafts allowed the operator to control the position of each arm independently. The regulator was generally set either vertically or horizontally, and the indicators in 45-degree increments; 92 different signals could be sent. Abraham Edelcrantz traded compactness for mechanical simplicity (*right*). His system had 10 "shutters" that could be set in either a vertical or horizontal position, for a total of 1,024 signals.

SWEDISH OPTICAL TELEGRAPH



per minute would arrive at each end of the line.

Because each signal could encode any of 25,392 words and phrases, two signals per minute corresponds to a data rate of a little less than 15 bits (which can encode 32,768 combinations) per 30 seconds, or 0.5 bit per second. If the average length of a word or phrase from the codebooks was about 10 characters, the effective speed would have been 20 characters per minute. Surprisingly, this speed compares very favorably with that of the first electromagnetic telegraphs. The "single-needle telegraph," patented by William Fothergill Cooke and Charles Wheatstone in 1837, averaged 25 characters per minute. The first system that was faster was the Wheatstone Automatic Telegraph of 1858, which used punched paper-tape readers to achieve speeds up to 2,000 characters per minute.

While Chappe was battling the Revolution to build a telegraph network in France, his counterpart a few hundred kilometers to the north was making his way in an atmosphere of assassination and court intrigue. Abraham Edelcrantz was born Abraham Niclas Clewberg on July 29, 1754, in the Swedish city of Åbo, now part of Finland. Like Chappe, he was bright and came from a relatively well-to-do family. Before his 19th birthday he had written two doctoral theses, one in optics and the other in literature. During the next few years, he taught

courses in both electricity and literature at the Royal Academy in Åbo. When King Gustav III visited in 1775, the young Clewberg was asked to recite some poetry and apparently made such an impression that he received an invitation to visit the court in Stockholm.

Clewberg moved to Stockholm permanently in April 1783. Within three years he had gained a prestigious membership in the Swedish Academy of Literature and Arts. In 1787 he became the king's private secretary, and two years after that, just 35 years old, he was raised to the peerage and changed his name to Edelcrantz. The assassination of Gustav III in 1792 (the plot of Verdi's 1859 opera *A Masked Ball*) started a period of relentless persecution of everyone who had been associated with the monarch. Like Chappe in similar circumstances, Edelcrantz somehow managed to avoid being targeted. Many of his friends were not so lucky and were thrown in prison or forced to leave the country.

In 1794, when news of the French telegraph spread rapidly through Europe, Edelcrantz was fascinated. Solid information was hard to come by; most accounts were based on scanty reports from travelers who had merely seen the French telegraph or read newspaper reports. Edelcrantz immediately set to work on his own system and demonstrated a prototype on November 1, 1794—King Gustav IV's 14th birthday.

Edelcrantz's first design was similar to Chappe's. It consisted of a single

support beam with two rotating indicators. Each of the indicators could be set in one of four distinct positions, 16 combinations in all. A reduced but still useful alphabet could be represented in semaphore positions.

The first test used three stations: one on the roof of the Royal Castle in the center of Stockholm, a second about five kilometers away on the outskirts of the city and the third seven kilometers farther on the grounds of the Royal Palace in Drottningholm. Within a week of the demonstration, the young king requested his Council of Advisers to study the construction of an optical telegraph network, with connections to Denmark and Finland. He also appointed Edelcrantz to the council, virtually sealing the outcome of the study.

On January 30, 1795, Edelcrantz started construction on the first telegraph line, from the Katarina Church in the center of Stockholm to the fortress of Vaxholm, about 35 kilometers away. The line went into operation on July 28. Between 1795 and 1797 two more lines were built: from Stockholm to Fredricsborg and from Grisslehamn to Signilskär and Eckerö on Åland. By this time Edelcrantz had abandoned the semaphore. His new telegraph consisted of a matrix of nine shutters with a tenth large one mounted on top. By opening or closing each shutter independently, an operator could convey 1,024 different signals.

In 1801 Swedish telegraphers built another, but ill-fated, line to Helsingborg near the Danish border. The station at Helsingborg was intended to establish a connection between the Swedish optical telegraph network and the beginnings of a Danish network. Three days after the link was opened, the British fleet attacked Denmark. Swedish commanders failed to respond to a telegraphic call for aid, and the British bombarded Copenhagen. The Danes understandably lost interest in the connection, and nothing further came of the line.

This early demonstration that the value of technology depends on the use people make of it did not dim Edelcrantz's fame. In 1796 he had documented his efforts in *A Treatise on Telegraphs*. The book was soon translated into German (and French) and earned him a membership in the Swedish Academy of Sciences. In the spring of 1808 a Royal Swedish Telegraph Institution was created, and Edelcrantz was appointed its first director.

As Chappe had done before him, Edelcrantz took this opportunity to revise his codes. The new version allowed



MOBILE TELEGRAPH found use during Napoleon Bonaparte's invasion of Russia in 1812. Abraham Chappe (one of Claude's brothers) designed it at Napoleon's request.

for 5,120 signals. Moreover, the 13 tables of the expanded system gave Edelcrantz room to cope with certain administrative issues—he added, for example, signals for the punishment of negligent operators.

The curious punishment of stepping onto the telegraph arms appears to have been popular at some locations later in the century—among them a station close to Göteborg. The author Nils Risberg notes that the operators in question were the station superintendent's daughters. Risberg interviewed the two in the 1930s and wrote that they remembered the experience well but called it “just fun.”

The new tables also brought with them a motto for the Telegraph Corps: signal 636 (*Passa väl upp*, or Be on Guard). The signal appeared prominently in the seal of the Telegraph Institution and on the buttons of the telegraph operators' uniforms, to serve as a permanent reminder of their duty.

In November 1809 the Swedish network consisted of approximately 50 stations spread out over a distance of some 200 kilometers and provided employment for 172 people. It included lines from Stockholm to the city of Gävle in the north, Landsort in the south and Eckerö on Åland in the east. Shortly thereafter, however, Gustav IV fought a disastrous war with Russia in which Sweden lost Finland and he the crown. The network was dismantled, not to be rebuilt for more than a generation. When Edelcrantz died in 1821, the towers he had constructed still lay in ruins.

It was not until the middle of 1836 that optical telegraph lines from Stockholm to Vaxholm and to Sandhamn came back into operation. By 1838 the net had been extended to a rough semblance of its former extent. The last addition to the optical telegraph network was made as late as 1854, when telegraphers extended the Furusund line to Arholma and Söderarm.

In 1840 almost every country in Europe had at least one or more optical telegraph lines in service. In England, between 1796 and 1816, the British Admiralty operated lines from London to Portsmouth, Plymouth, Yarmouth and Deal. In Germany, a line ran from Berlin through Potsdam and Magdeburg via Köln and Bonn to Koblenz, starting in 1832. Others carried messages from Hamburg to Altona and Cuxhaven and from Bremen to Bremerhaven. Russia entered the race relatively late but in grand style, with the opening of a line of 220 semaphore stations on April 8, 1839, between St. Petersburg and Warsaw. Optical telegraph lines also spanned parts of the U.S.



TELEGRAPH OPERATOR'S UNIFORM was approved by Swedish king Carl XIV Johan in 1809. The uniform button (*inset*) shows the motto of the Telegraph Institution, signal 636: Be on Guard.

where the electrical cables at first could not. In 1864 there were 174 electromagnetic telegraph stations, with 250 operators, as well as 24 optical telegraph stations manned by 66 operators. Not until 1881 were the last three optical telegraphs replaced in Sweden, bringing their era to a close.

Claude Chappe and Abraham Edelcrantz's achievements are mostly forgotten, superseded by developments they could not have imagined in their most ambitious moments. They were, nonetheless, the true pioneers of data networking. Both of them had to solve many subtle problems to enable operators to transfer messages smoothly over long chains of stations. In retrospect, it is exceptionally interesting to see that some of their ideas have been rediscovered only recently by the designers of modern digital protocols. They not only conceived sophisticated methods for data compaction, error recovery, flow control and even encryption, they also put them to practice.

What characterized these two inventors was not luck but a strong vision and a relentless dedication to their goal, even under adversity. Technical obstacles were among the easier ones they had to overcome. As Edelcrantz wrote in his *Treatise*: “It often happens, with regard to new inventions, that one part of the general public finds them useless and another part considers them to be impossible. When it becomes clear that the possibility and the usefulness can no longer be denied, most agree that the whole thing was fairly easy to discover and that they knew about it all along.”

Yet even as optical telegraph networks reached their zenith, their electromagnetic rival was beginning to make inroads. In 1837 England and the U.S. began the move from tower and telescope to copper wire and code key.

To the countries with established optical networks, it was not immediately clear that the change would be an improvement. The French optical telegraph network, for example, proved immune to change for almost 10 years. The first electromagnetic telegraph replaced the historic line from Paris to Lille in 1846. The initial signals were passed by means of a curious device, designed by Alphonse Foy and Abraham Louis Breguet, that reproduced the positions of the Chappe semaphore. The optical network reached its peak in 1852, with 556 stations along 4,800 kilometers of lines. It linked 29 of France's largest cities to Paris. The administration employed as many as six operators per station, working in shifts—more than 3,000 workers in all.

In Sweden, replacement of the optical telegraphs began even later. For more than 10 years, electrical and optical telegraph stations were in use side by side; the optical telegraphs reached

FURTHER READING

- PIONEERS OF ELECTRICAL COMMUNICATION. Rollo Appleyard. Macmillan & Company, 1930.
- A HISTORY OF TACTICAL COMMUNICATION TECHNIQUES. David L. Wood. Ayer Company Publications, 1974.
- THE OLD TELEGRAPHS. George Wilson. Phillimore Chichester, 1976.
- LA TÉLÉGRAPHIE CHAPPE. Fédération Nationale des Associations de Personnel des Postes et Télécommunications pour la Recherche Historique. Editions de l'Est, Nancy, 1993.
- THE EARLY HISTORY OF DATA NETWORKS. Gerard J. Holzmann and Björn Pehrson (in press).

A War Not Won

by Tim Beardsley, *staff writer*

There was so much good news at their meeting in September that the members of the President's Cancer Panel might have been pardoned had they been overwhelmed by euphoria. Reports of promising therapies and diagnostic maneuvers swirled with accounts of deep new insights into the underlying genetics and molecular biology of the disease. Then the mood grew somber—or was it tense? John C. Bailar III, a noted professor of epidemiology and biostatistics at McGill University, began his observations on recent trends in the morbidity and mortality of cancer.

Bailar had created a storm in 1986 after publishing a damning lack-of-progress report on the "war on cancer" initiated by President Richard M. Nixon when the chief executive signed the National Cancer Act in 1971. Bailar's unflinching summary of the latest body counts last fall, part of a formal evaluation of the national cancer program, led to the same disturbing result. "In the end, any claim of major success against cancer must be reconciled with this figure," he said, pointing to a simple graph that showed a stark continuing increase in U.S. death rates from cancer between 1950 and 1990. "I do not think such reconciliation is possible and again conclude, as I did seven years ago, that our decades of war against cancer have been a qualified failure. Thank you."

The numbers, based on data supplied by the National Cancer Institute (NCI), indeed present a grim picture. Bailar's principal conclusion, with which the NCI agrees, is that U.S. cancer death rates went up by 7 percent between 1975 and 1990. This number, like all those Bailar cited, has been adjusted to compensate for the changing size and composition of the population with respect to age, so the increase cannot be blamed on people's dying less often from other diseases. Cancer is the second leading cause of death in the U.S. after heart disease; 526,000 cancer deaths were expected to have occurred in the U.S. in 1993.

What do these facts mean? Would analysis

of them reveal environmental or other influences that are triggering the disease? Do the results justify the massive, frontal assault, which relies on research into the fundamental causes, on efforts to devise sensitive diagnostic procedures and on empirical attempts to develop cures? Have the researchers and clinicians been barking up the wrong trees for the past two decades?

Anyone trying to answer these questions by probing more deeply into the data finds that the devil of uncertainty lurks in the details. At first, the basic facts seem simple. By far the greatest contribution to the overall increase comes from lung cancer, which accounts for more than a quarter of all cancer deaths. Most lung cancer deaths are, researchers agree, the result of cigarette smoking. Although deaths from lung cancer have started to fall in men, they have more than doubled in women since 1973, because women took up smoking decades after most men did. Today lung cancer kills even more women in the U.S. than does breast cancer.

If lung cancer is excluded—an exclusion that the blunt-spoken Bailar dislikes because self-inflicted deaths should be as much a matter of concern as any other—the total cancer death rate has changed little since the war on cancer was first waged. Yet the total conceals a variety of individual cancers that apparently have increasing death rates. They include non-Hodgkin's lymphoma, multiple myeloma and cancers of the prostate, brain, kidney, esophagus and breast. All these increases are statistically significant by Bailar's reckoning, even though the biggest of them is only a tenth the size of the explosion in lung cancer deaths.

Thus, the war against cancer is one in which the foe's order of battle changes con-

BRAIN SCAN in this patient is made by means of magnetic resonance imaging. Sophisticated new technology eases the task of detecting tumors. But such advances also complicate the interpretation of trends in the occurrence of disease by uncovering cases that clinicians using less powerful methods might have missed.

*Despite dramatic scientific
gains, cancer remains
an undaunted killer*



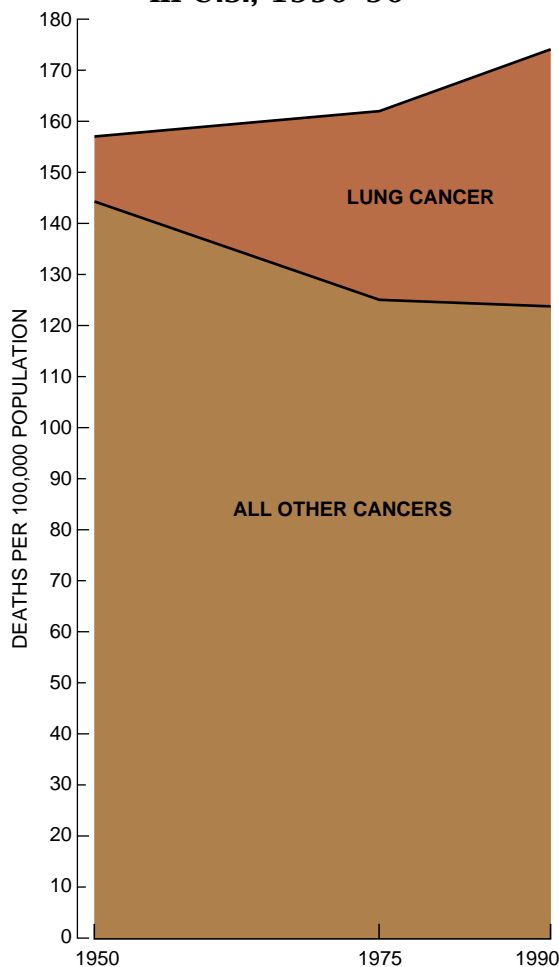
stantly. Death rates have declined for such cancers as those of the colon and rectum, stomach, uterus (including the cervix), bladder, cranium, bone, gallbladder and testis. Death rates from cancers in children fell by almost half between 1973 and 1989, in large part because of better treatments. Yet because cancer in children was rare to begin with, this improvement—and smaller gains in young adults—has had only a minuscule effect on the big picture. Overall, the increases in cancer death rates are twice as large as the decreases.

Another measure of the burden of cancer is its incidence—that is, the number of people in the population who are recorded as developing the disease. The President's Cancer Panel, chaired by Harold P. Freeman of Harlem Hospital Center, heard those figures from Edward J. Sondik, a senior NCI statistician. They, like the death rates, are adjusted for the changing composition and size of the population. They show first and foremost an increase of more than 100 percent in lung cancer in women between 1973 and 1990. Melanoma and prostate cancer also displayed very large increases (of more than 80 percent) during that period. The rogues' gallery of seemingly increased incidence also includes non-Hodgkin's lymphoma, multiple myeloma and cancers of the breast, kidney, liver and brain.

Cancers apparently decreasing in incidence include those of the cervix, uterus, stomach, pancreas and mouth; Hodgkin's disease and some leukemias also belong in this category. Sondik concluded that the overall incidence of cancer had increased by 18 percent between 1973 and 1990. The NCI says some childhood cancers are becoming more common. They include acute lymphocytic leukemia as well as cancers of the brain and nervous system.

Does this picture suggest changes in underlying causes, or are the variations in incidence artifacts of the development of medical knowledge? Diagnostic practices and technology have become more sophisticated and so could account for some of the trends in both mortality and incidence, experts argue. The widespread availability of nuclear magnetic resonance imaging, for example, means that today it is unlikely someone dying of brain cancer will be misdiagnosed as succumbing to a

Annual Cancer Deaths in U.S., 1950–90



SOURCE: John C. Bailar III

CASUALTY REPORT from the war on cancer shows that the effort has not slowed deaths from the disease in the U.S., where mortality from cancer is increasing even if (as here) the numbers are adjusted to allow for the aging of the population.

stroke. Twenty years ago that may have been a common occurrence. The increase in deaths from brain cancer might, therefore, be attributable at least in part to better diagnosis.

Incidence data are the most susceptible to distortion. Bailar is so skeptical of them that he declined to discuss the topic in September at the President's Cancer Panel meeting. He would only express his conviction that the reported significant increases in incidence of lung and prostate cancer are "largely or entirely spurious." In Bailar's view, improving medical techniques now allow the identification as "cancer" of many tumors that would otherwise never have become noticeable. "Everybody is terrified of missing a diagnosis," he told SCIENTIFIC AMERICAN.

The increase in the incidence of breast cancer is subject to such distor-

tion. The recorded incidence of the illness grew markedly during the mid-1970s after First Lady Betty Ford and Margaretta (Happy) Rockefeller, wife of Vice President Nelson A. Rockefeller, were diagnosed as having the disease. Apparently, the highly visible cases of the illness prompted many women to be screened. Still, Peter Greenwald, head of the division of cancer prevention and control at the NCI, says there has been a steady, slow increase in the incidence of breast cancer since the 1970s. On the other hand, he believes a seeming acceleration in the 1980s was related to the continuing increase in demand for mammography screening. Improvements in surveillance can, paradoxically, also precipitate a decrease in incidence. Screening by Pap smears, which detect treatable precancerous changes, is thought to have led to a 35 percent drop in the reported occurrence of cancer of the cervix.

Despite the difficulties of interpretation, epidemiologists widely read the incidence data for clues to causes of cancer. If a particular type of cancer is truly becoming more common, there is presumably some identifiable and perhaps modifiable etiology. Devra Lee Davis, now a senior adviser to Philip R. Lee, the assistant secretary for health, has created a furor in epidemiologic circles in recent years by contending that many of the increases in cancer incidence as well as in death rates are not merely clinical artifacts.

Davis points by way of example to a recent study in Canada. The investigator, Marie Desmeules of Health and Welfare Canada, assessed the effect of computed tomography and magnetic resonance imaging on detection of brain cancer and other neurologic diseases. She arranged for suspected cases of brain cancer to be reclassified after she had removed from patients' charts evidence acquired with these techniques. The results suggested that in Canada those instruments were partly but not entirely responsible for the twofold increase in brain cancer diagnoses among the elderly.

Changing diagnostic practices are thought to have less of a distorting effect on cancer death rates than on incidence rates. Even so, the interpretation of death rates is also controversial. In principle, death rates should reflect both improvements in treatment and changes in incidence. Unfortunately, co-

lorectal cancer is the only one of the big four killers—lung, colorectal, breast and prostate—that is becoming more curable. By a cruel twist of fate, the other cancers that can now be cured somewhat more successfully than in 1973 are relatively rare. They are Hodgkin's disease, some leukemias, cancer of the thyroid, testicular cancer and, perhaps, cancers of the uterus and bladder. (Earlier detection probably explains some of the improvements in survival after diagnosis.)

Some cancers in which death rates are decreasing probably cannot be chalked up as victories. Declines in deaths from stomach cancer and cancer of the uterus, mainly the cervix, are global and started decades ago, according to Bailar. They therefore probably owe little to advances in therapy. Bailar agrees, however, that Pap smears have helped reduce the number of deaths attributable to cervical cancer. Better food preparation is commonly held to account for the decline in stomach cancer. Yet no one has certain knowledge of the reason (or rather, as Bailar dryly observes, many people know, but they cannot agree).

The picture becomes even more complicated when the data are sliced in different ways. One cut bares the rates of death and disease at different

ages. Davis, Sondik and David G. Hoel of the Medical University of South Carolina and other experts maintain that consideration of such isolated rates can reveal important patterns. Specifically, the death and incidence rates for some cancers are staying roughly constant or decreasing in young people while increasing in older people.

Breast cancer deaths, for example, conform to this pattern. They have decreased in women younger than 50 years but have become more common in women above that age. This is happening even though the total breast cancer death rate, according to Greenwald of the NCI, may have just started to decline. The number of deaths from melanoma, and the incidence of this highly lethal cancer, which first emerges on the skin, has decreased in people younger than 45 while increasing in older individuals.

In six regions covering 15 industrialized countries (including the U.S.), Davis and her colleagues found that over a period of 17 years, mortality from lung, breast and prostate cancer increased in individuals older than 45. Deaths from brain and other nervous system cancers among people older than 75 show "drastic" increases in several major industrial countries. And in the U.S., incidence of brain cancer

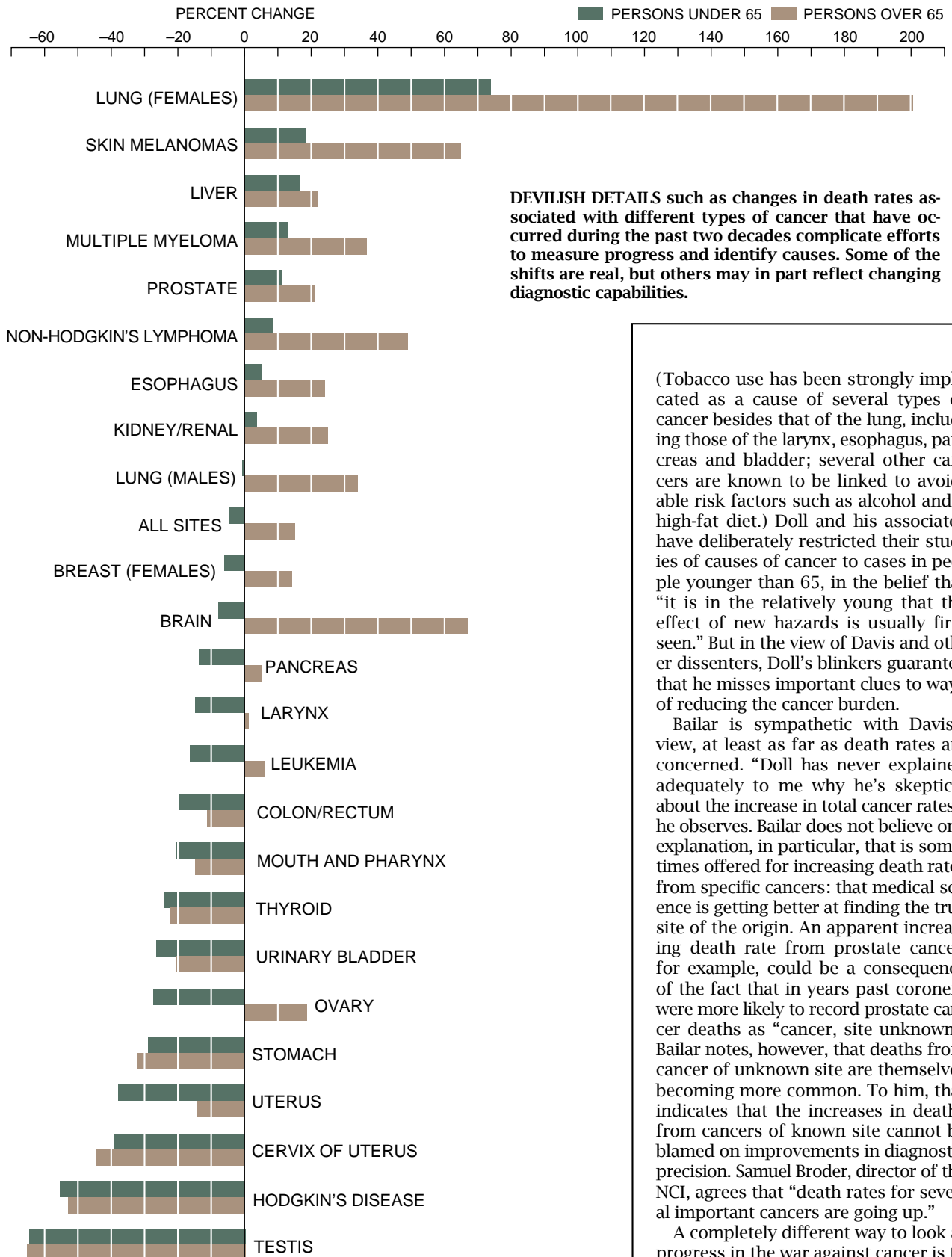
seems to have increased among young people. These trends, referring as they do to rates rather than absolute numbers, again cannot be explained by the increasing average age of the population. Old people are definitely more likely to be diagnosed with and recorded as dying from cancer than they were some 20 years ago. But Davis's vigorous contention that these trends manifest new, possibly environmental threats to public health has provoked skepticism.

The prominent English epidemiologist who first identified the link between smoking and lung cancer, Sir Richard Doll, has been one of Davis's chief critics. While not disputing that recorded death rates for some cancers have risen among the middle-aged and elderly, Doll and others of like mind, such as Brian E. Henderson of the Salk Institute for Biological Studies in San Diego, believe almost all the increases can be attributed to tobacco use, to other avoidable exposures such as sunlight or alcohol, or to diagnostic artifacts.

DECLARATION OF WAR on cancer is made as President Richard M. Nixon signs the National Cancer Act on December 23, 1971. The legislation launched the national cancer program, which has since expended some \$25 billion.



Changes in U.S. Cancer Death Rates,* 1973-90



DEVILISH DETAILS such as changes in death rates associated with different types of cancer that have occurred during the past two decades complicate efforts to measure progress and identify causes. Some of the shifts are real, but others may in part reflect changing diagnostic capabilities.

(Tobacco use has been strongly implicated as a cause of several types of cancer besides that of the lung, including those of the larynx, esophagus, pancreas and bladder; several other cancers are known to be linked to avoidable risk factors such as alcohol and a high-fat diet.) Doll and his associates have deliberately restricted their studies of causes of cancer to cases in people younger than 65, in the belief that "it is in the relatively young that the effect of new hazards is usually first seen." But in the view of Davis and other dissenters, Doll's blinkers guarantee that he misses important clues to ways of reducing the cancer burden.

Bailar is sympathetic with Davis's view, at least as far as death rates are concerned. "Doll has never explained adequately to me why he's skeptical about the increase in total cancer rates," he observes. Bailar does not believe one explanation, in particular, that is sometimes offered for increasing death rates from specific cancers: that medical science is getting better at finding the true site of the origin. An apparent increasing death rate from prostate cancer, for example, could be a consequence of the fact that in years past coroners were more likely to record prostate cancer deaths as "cancer, site unknown." Bailar notes, however, that deaths from cancer of unknown site are themselves becoming more common. To him, that indicates that the increases in deaths from cancers of known site cannot be blamed on improvements in diagnostic precision. Samuel Broder, director of the NCI, agrees that "death rates for several important cancers are going up."

A completely different way to look at progress in the war against cancer is to examine what proportion of people diagnosed with different types survive for at least five years. Although five-year

*age-adjusted

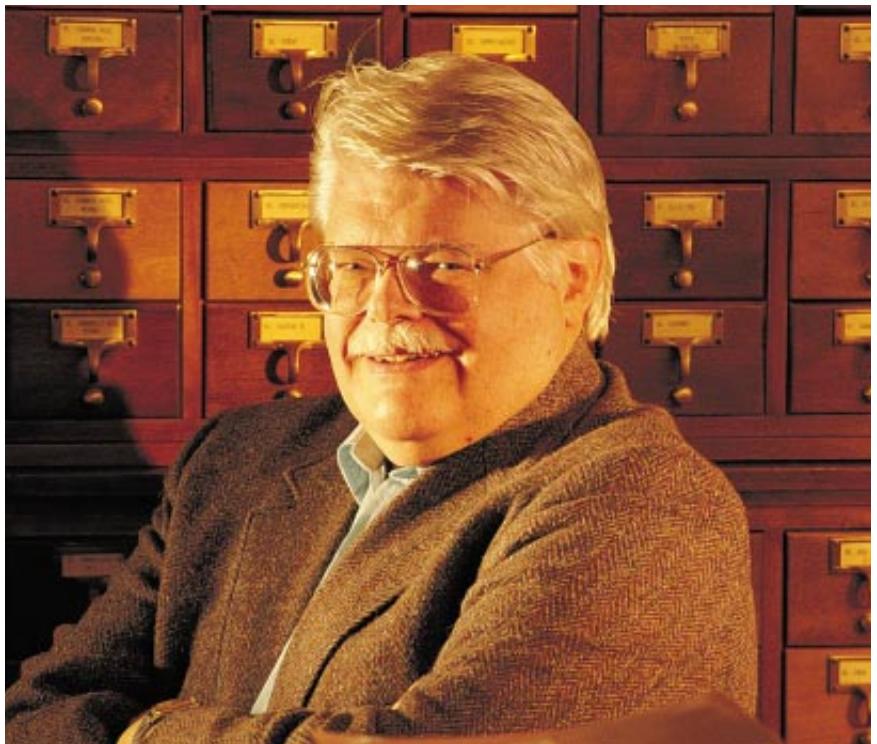
SOURCE: National Cancer Institute

CRITIC John C. Bailar III, a Canadian epidemiologist, has vetted the national cancer program consistently and sharply. "Whatever we have been doing," Bailar says, "it has not dealt with the broadly rising trend in mortality."

survival rates say nothing about incidence, they are seen as highly relevant by the people diagnosed as having cancer. Unfortunately, the comfort that these data offer is distinctly cold. Despite the heartening gains in cure rates in the young, the NCI estimates the overall improvement in five-year survival since Nixon launched the program is only 4 percent. Bailar told the President's Cancer Panel that even that figure may constitute an overestimate.

The seeming impotence of the NCI in the face of what is still one of the most feared diseases has led a vocal chorus of critics to argue that the organization should change its tack. With a budget of more than \$2 billion this year, the NCI is the largest of the national health institutes by a substantial margin. Second in line is the National Heart, Lung and Blood Institute, which can point to a 30 percent reduction in deaths from heart disease since 1975 as evidence of its success in pioneering treatments and, significantly, in promoting exercise and a healthy diet. The NCI can report no victories on that scale. The institute was mandated by the National Cancer Act to "comprehensively and energetically exploit" scientific leads to combat the disease. But the critics accuse it of having neglected research aimed at prevention in favor of the search for cures. There has been little to show for the \$25 billion spent on the war on cancer, they charge: oncologists still mainly cut, poison, burn and hope.

To be sure, not all progress is reflected in death and incidence rates. The President's Cancer Panel heard in September how improvements in surgery mean patients are now more likely to retain breasts, limbs and bowels. New therapeutic agents such as taxol, a derivative of Pacific yew tree bark, crop up from time to time. Side effects of chemotherapy and radiotherapy such as nausea are also managed better than they once were. And the panel heard much about the numerous advances in understanding cancer on the molecular level. Oncogenes have been found that promote cancer when they are damaged, tumor suppressor genes have been identified that prevent cancer as long as they are present, and mutations have been charted as they accumulate and eventually send a healthy cell over the brink to a cancerous state.



Optimists, such as Greenwald, feel that large-scale improvements in death rates and incidence rates will eventually result from the years of investment in basic cancer research, although Greenwald guesses it may be a decade before the benefits are widely felt. Gene therapy, immunotherapy and antisense RNA technology are just a few of the methods now in early testing. Such subcellular interventions could, according to their proponents, bring about dramatic gains by changing the activity of specific genes in tumors and selectively modifying the immune response. But skeptics, including Bailar, say they have heard that kind of talk too many times before. "No knowledgeable person can continue to believe there is necessarily a spectrum of marvelous cures of cancer waiting to be found," he asserts. Bailar says he is fed up with the "constant procession of hopeful news stories" suggesting a cure is just around the corner.

Existing chemotherapies, despite improvements, are still arduous. Oncologists who employ them know they are wielding double-edged swords. Some of the treatments for lymphoma and leukemia trigger other cancers, after therapy for the initial disease has been successfully completed.

Unrealistic announcements by the NCI have probably only added to the critics' impatience. In 1984, under a previous director, Vincent T. DeVita, Jr., the NCI announced with fanfare the "achievable" goal of cutting cancer

deaths in half (from 1980 levels) by the year 2000. The institute has since stayed quiet about the fact that the goal has been getting further away with each passing year. Still, Broder is diplomatic about how the NCI has changed. He defends the principle of establishing targets, pointing out that the state of Utah has (almost) achieved the goal for one cancer, that of the lung, by reducing smoking. But he acknowledges that the 50 percent goal will not be attained for the nation as a whole. "If there has been a change," says Broder, emphasizing the "if," "it is that we must shed our illusions that cancer is an easy problem. It is a formidable problem."

Bailar is among those who believe the NCI should devote much more of its money to prevention. "We've come to the point where we must face a really serious problem square in the face. What if there aren't any major advances to be obtained in chemotherapy?" he demands. "For a lot of years now, we've been tinkering. It's not going to solve the big problem of cancer, and we need a major advance." Prevention "is going to involve everybody over their whole lives," he declares. "It's going to involve cleaning up the workplace and the environment, it's going to involve changing our diets, and it's certainly going to be a bigger hassle and more expensive than our ideal treatments would be."

The NCI is already nominally devoting more of its budget to prevention research than to treatment research. But Broder admits that it is hard to classify



MAVERICK Devra Lee Davis, an adviser at the Department of Health and Human Services, proposes that synthetic chemicals permeating the environment and mimicking estrogen in the body could be causing breast cancer.

research unambiguously as being related either to treatment or prevention. "Take biological carcinogenesis—is that treatment or prevention related?" he asks. (He refers to cancer caused by viruses and classed as prevention related.) The agency is, for example, pursuing studies in which nutrients or drugs are given to populations at risk in the hope of forestalling the development of cancer. Late last year NCI scientists established that beta-carotene, in combination with vitamin E and selenium, reduced deaths from cancer in a trial in Linxian, China. In the U.S., the NCI is studying drugs that might achieve the same effect.

Ironically, these studies have engendered fierce controversy. Opponents are attacking a breast cancer prevention study under way with tamoxifen because the substance is known to cause cancers of the liver and endometrium. Researchers hope and expect that tamoxifen will prevent more illness than it causes. Greenwald defends the study as "one of the most important prevention trials we can do." He points out that there is already evidence that tamoxifen can prevent breast cancer. But he predicts that "we'll probably face the same issues with the finasteride trial," referring to a prostate cancer prevention trial that started recently.

Indeed, a quality of blindman's buff suffuses efforts to lay the groundwork for preventive therapies. Researchers have made dramatic progress in the effort to unravel oncogenesis. The discovery of oncogenes and antioncogenes, as well as the identification of a new gene that triggers colon cancer by causing mutations, comes to mind. Yet understanding of the genetic roots of can-

cer, Greenwald acknowledges, is for the present far from complete.

The uncertainty that haunts investigators who seek preventable causes follows the workers into epidemiology as well. Strong evidence proves that tobacco leads to about a third of cancer deaths. The magnitude of the toll exacted by other causes of cancer is unclear. Among them are diet, alcohol consumption, several industrial chemicals (asbestos, for one), several viruses (some of them sexually transmitted), and radiation (most though perhaps not all of it from natural sources).

The NCI estimates that about a third of cancers could be brought on by diet. Greenwald says that figure could lie anywhere between 20 and 60 percent. The principal evidence concerning humans, which is not conclusive, consists of studies that examine the rates of cancer in people who emigrate. Breast cancer in Japan, for example, is four times lower than it is in the U.S. When Japanese women move to the U.S., they or their daughters eventually acquire the disease at North American rates. A diet rich in fat is often held to explain such observations. The data supporting that conclusion are weak, however, and are based mainly on animal studies. Other studies in people have produced conflicting results. "What causes breast cancer? I don't think anyone knows," Broder observes. In another component of its inquiry into possible dietary causes, the NCI is now studying carcinogenic chemicals formed in meat cooked at high temperatures.

Actually, for some cases of breast cancer, there is a good explanation: a gene called *BRCA1*. Researchers have been hunting for this gene, and others

that probably provoke some breast cancers, for years. By all accounts they are now close. In many types of cancer, a proportion of cases seem to be inherited. Greenwald speculates that when tests for *BRCA1* and other cancer genes become widely available, environmental and hereditary cancer risks will be managed as heart disease risk is managed today. If so, basic research would pay off by enhancing the effectiveness of preventive measures. Colon cancer susceptibility is already being managed better because genes are recognized that are involved in some cases.

In the view of many of the critics, however, the evidence that cancer is becoming more common should stimulate a search for unknown environmental factors. Industrial chemicals are naturally suspect. "The increased volume and the diversity of synthetic chemicals manufactured since World War II raise serious concern about the cumulative population risks of exposure to these substances," asserts Philip J. Landrigan of the Mount Sinai School of Medicine, an expert on occupationally caused cancer.

David P. Rall, a former director of the National Institute of Environmental Health Sciences, told the President's Cancer Panel that "much better data are needed on ambient levels of pollutants." Most of the chemicals people are exposed to, he said, have not been adequately tested. A conviction that the NCI is downplaying the importance of pollution led Samuel S. Epstein of the University of Illinois, together with 72 co-signatories, to lambaste the NCI last year. The group accused the "cancer establishment" of having "continually minimized the evidence for increasing cancer rates, which it has largely attributed to smoking and dietary fat, while discounting or ignoring the causal role of avoidable exposures to industrial carcinogens in the air, food, water and the workplace."

The notion that industrial carcinogens might be more important than has been recognized by the NCI seems to have gained ground recently. Davis, in collaboration with Aaron E. Blair, chief of occupational studies at the NCI, has found that many studies indicate that farmers have higher rates of specific cancers than do other people, even though in other respects they are

healthier. The types of cancer that are more common in farmers include Hodgkin's disease, multiple myeloma, leukemia, melanoma and cancers of the lip, skin and prostate. Several of these, Davis points out, are among those becoming more common in the population at large, thus adding further weight to the idea that pollutants such as pesticides might be to blame. Other factors could well be implicated, however, including exposure to sun, animal viruses and vehicle exhausts.

To settle the question, the NCI is planning a collaborative study with the Environmental Protection Agency that will enroll more than 100,000 farmers and their spouses. Greenwald says the NCI at present believes no more than 8 percent of cancers (probably about 5 percent) are caused by industrial pollutants. But he acknowledges that the figure could be higher. Landrigan estimates that about 10 percent of cancers are caused by occupational exposures to carcinogens. The epidemiologist believes a further 5 percent is attributable to pollution by artificial chemicals.

Davis and others have developed a specific hypothesis that they believe might explain the increasing rates of breast cancer, in particular. Their proposition rests on the fact that risk for the disease is linked to exposure to estrogen. So they have lit a fire under the cancer institute by proposing that the rise in recorded incidence and in deaths among older women might be explained by increasing environmental levels of fat-soluble synthetic chemicals that mimic or amplify the physiological effects of estrogen. Some such chemicals are known to be carcinogenic from laboratory studies. Moreover, several plastics that are in widespread use in laboratories have been shown to shed chemicals with estrogenlike properties in amounts sufficient to have hormonal activity on cultured cells.

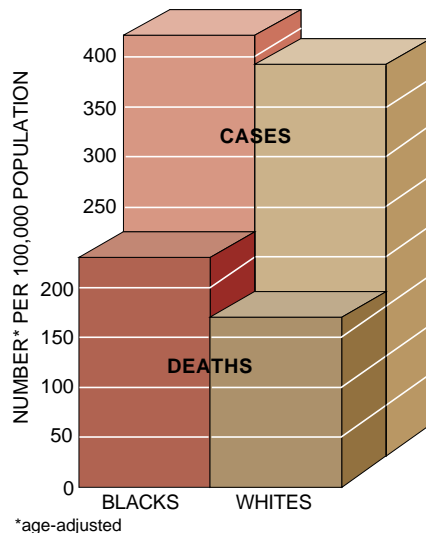
The so-called xenoestrogen hypothesis remains unproved. Nevertheless, Mary S. Wolff, a researcher at the Mount Sinai School of Medicine, conducted an experiment last year that offers support for it. She showed that women with breast cancer have higher than normal blood levels of DDE, an estrogenic metabolite of the now banned organochlorine pesticide DDT. Xenoestrogen is "definitely being taken more seriously" at the NCI, comments Louise A. Brinton, chief of environmental studies. The institute is, for example, planning to establish a laboratory of hormonal carcinogenesis. Others are doubtful, and some are frankly skeptical. The xenoestrogen idea "hasn't gained in credibility," notes Salk Institute researcher

DEADLY DISCRIMINATION is suffered by African-Americans, for whom cancer rates are higher than for whites. Relative lack of access to health care may explain some of the discrepancy in mortality and incidence.

Henderson. "The people who are pushing this feel they want to make a contribution, and they are simply responding to public fear by trying to identify a removable cause."

Despite the new attention the NCI is giving to environmental causes, and an increasing budget for prevention in general, Broder will not countenance the kind of major change in strategy that Bailar and Epstein are advocating. He argues that many of the institute's achievements in understanding the molecular biology of cancer have proved valuable in combating other diseases, AIDS in particular. And his response to the agency's critics is that the NCI "cannot directly influence the practice of care." Epstein, he asserts, "in effect wants us to be the Environmental Protection Agency or the National Institute of Environmental Health Sciences. We follow the advice of peer review, but

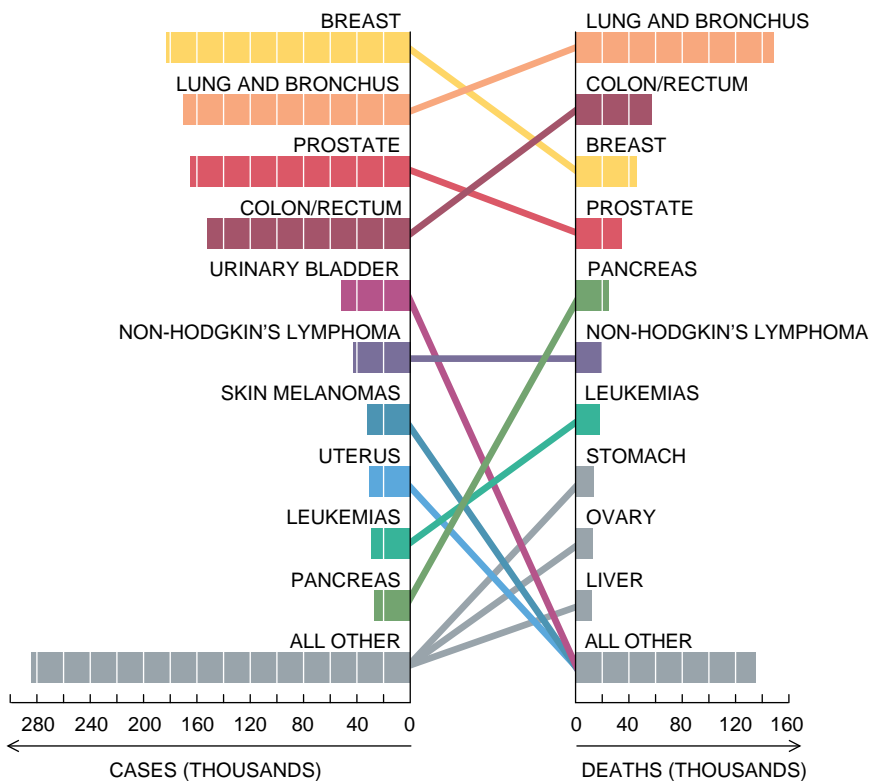
Relative Impact of Cancer in U.S. by Race, 1990



our primary mission is to generate knowledge."

There is still plenty of knowledge that needs to be generated. High among Broder's priorities is learning the reason for the variations between racial groups. Blacks in the U.S. die from can-

U.S. Cancer Cases and Deaths in 1993*



SOURCE: National Cancer Institute

FOUR KILLERS, cancers of the lung, breast, bowel and prostate, account for more than half of the cancer deaths in the U.S. The past 20 years have seen little improvement in cure rates for these major scourges.



LEADER OF THE BATTLE Samuel Broder, director of the National Cancer Institute, says the public must understand that basic research is needed if modest progress is to be extended. "Cancer causes suffering beyond metaphor," he declares.

cer 35 percent more frequently than do whites, although the incidence in that population is only 8 percent higher. Moreover, "colorectal cancer deaths are increasing in African-Americans while they are decreasing in whites. Something is going on here. It may have to do with access to care, but we need to find out," he asserts.

A key obstacle to gathering the knowledge that emerged at the President's Cancer Panel meeting was the expense of running clinical trials. At present, fewer than 3 percent of all cancer patients enter experimental protocols, whereas 80 percent of children with cancer do so. "I think the challenge of the future is to increase the number of patients who go into clinical trials," Paul Carbone of the University of Wisconsin told the panel. Carbone, an expert on cancer treatment, has chaired a special congressionally mandated review of measures of progress against cancer. Prevention trials cost less per volunteer but are still expensive because they enroll thousands of participants.

Moreover, prevention trials entail extraordinary practical difficulties. A facet of the Women's Health Initiative, a \$625-million, 14-year study financed by the National Institutes of Health, illustrates the problem. The study, which is now under way and aims to recruit 160,000 women older than 50, seeks in part to determine whether a reduction in dietary fat will lower the risk of breast and colorectal cancer. The Institute of Medicine recently reviewed the study and judged that as currently planned it is incapable of answering that question.

Those who conduct a prospective study of behavior, the institute noted,

have trouble guaranteeing that the subjects do what they are supposed to do. Individuals participating in a long trial have difficulty sustaining substantial changes in diet. Another difficulty, the institute determined, is that participants who are told to reduce their fat intake are likely to compensate by consuming other sources of calories. Such behavior would blur the results by introducing other dietary factors that might be responsible for any observed effect on the incidence of breast and colorectal cancer among the participants.

Prevention trials that merely require people to take one or more dietary supplements are more likely to achieve good compliance, notes David J. Hunter of the Harvard School of Public Health. But even those trials have problems. One is that people in control groups, whose behavior is supposed to be unchanged by the intervention under study, often decide to emulate the intervention themselves. Thus, members of the control group in a major heart disease trial some years ago started to exercise like the intervention group. As a result, the difference between the groups was small.

Help may come from an unexpected quarter: the president's planned health care reforms. Many patients participating in trials now have difficulty obtaining reimbursement for their care. Broder says he expects that whatever reforms are eventually adopted will make it easier to conduct peer-reviewed clinical trials. He even speculates that universal coverage could lead to a standardized system for recording treatments and outcomes so that "in effect, everybody is in a large, simple trial."

Furthermore, findings of his cancer

panel are reported directly to President Clinton, so the chief executive may become familiar with the dilemmas of cancer research. The panel's findings are likely also to feed into the deliberations of the National Cancer Advisory Board, which is charting a future course for the much-criticized cancer institute. Broder insists that he is ready to respond to such scrutiny. "The public and scientists alike have to understand we need a basic research agenda," he states. "It's going to take a lot of effort. We need a strong clinical trials program, but we are making progress."

Although the institute's budget might appear large to an outsider, Broder defensively points out that it is small in comparison to the budget for the National Aeronautics and Space Administration (almost \$15 billion this year). Since 1980 the NCI budget has grown a mere 1 percent in real terms. "We need to make sure," Broder emphasizes, "that we always understand both the good news and the bad news."

So, taking the good news with the bad, what strategy should guide the war's conduct? In the view of John Cairns, an emeritus professor at the Harvard School of Public Health, the facts of economic life favor a preventive rather than a therapeutic emphasis in the long run. Apart from the indifferent results often achieved, therapy is expensive, Cairns notes. An intensive course of treatment for leukemia, he says, can cost more than \$100,000. Prevention could save lives more surely, at much lower medical cost.

Yet the knowledge needed to prevent—or treat—most cancers remains hidden in the recesses of the human genome and in the infinite subtleties of epidemiologic data. So implementation of any strategy waits on deeper understanding. Once again, it would seem, science must have the last word—and the last dollar.

FURTHER READING

EVALUATING THE NATIONAL CANCER PROGRAM: AN ONGOING PROCESS. President's Cancer Panel meeting, September 22, 1993. National Cancer Institute, Bethesda, Md., 1994. Available free by mail request.



RoboTuna

Seaborg may show the way to hydrodynamic efficiency

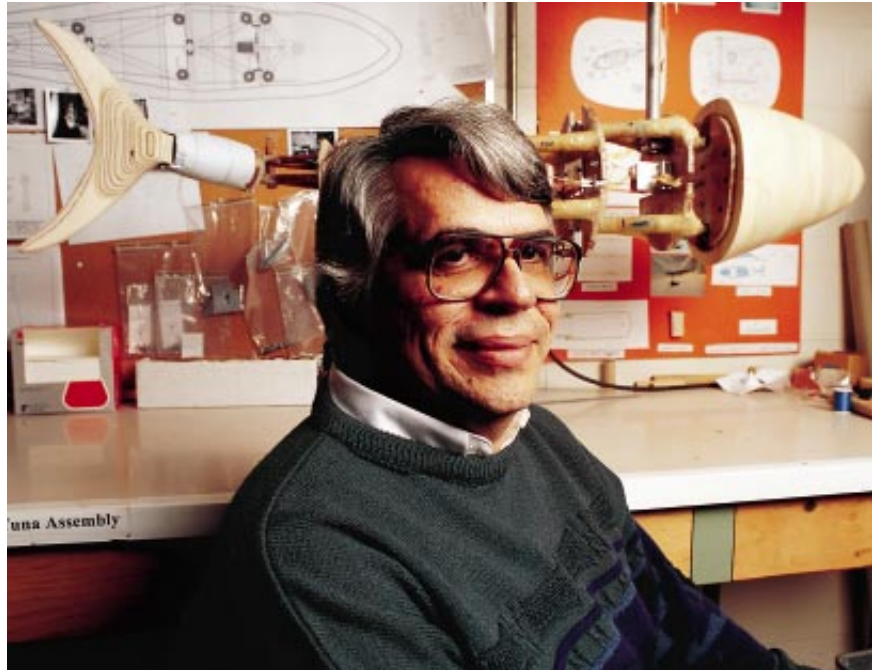
Anyone who has looked at the sleek, compact body of a tuna with its scimitar fins and seamless gills recognizes that it is a supreme piece of hydrodynamic engineering. A team of researchers assembled at the Massachusetts Institute of Technology has plans to capitalize on that fact by building and testing a mechanical chicken of the sea. They hope to discover ways to improve the efficiency of ships and submarines, open new avenues of research in fluid dynamics and perhaps even produce better designs for airplane wings.

The M.I.T. robotic tuna, or seaborg, as it might be called, is propelled by an articulated tail section, a fin that sweeps back and forth in the plane of the fish's movement. In addition to its tail, the five-foot-long seaborg consists of a wood skeleton with aluminum joints, all sheathed in a plastic skin. It was to be tested in December.

Unlike the real thing, the seaborg swims in a confined area, at M.I.T.'s Ship Model Towing Tank. An electric motor at poolside moves puppetlike strings that make the body undulate like a fish. Force and motion sensors on the mechanical beast track its movements. Michael S. Triantafyllou, an M.I.T. professor of ocean engineering and the father of the seaborg, is not under contract to Walt Disney Studios or to Steven Spielberg. His angels are the Advanced Research Projects Agency and the Office of Naval Research.

Triantafyllou works with his brother, George, a professor of mechanical engineering at City College of New York, and other investigators at both M.I.T. and the Woods Hole Oceanographic Institution. Anyone who is moved by this project to revive former Wisconsin senator William Proxmire's Golden Fleece Award should note that some fieldworkers have reported tunas attaining speeds of 20 nautical miles per hour, comparable to that of a commercial vessel. But fish expend a fraction of the energy of a fuel-guzzling boat.

The tail movement of a fish leaves a wake of swirling vortices that resemble the eddies produced by sticking a mo-



JESSICA BOVATT

SEABORG was being readied at M.I.T. for its first swim in December. Michael S. Triantafyllou designed the robotic tuna to gain a better grasp of how fish swim.

tionless paddle or hand into the water beside a moving boat. The wake from a paddle or hand creates drag that slows the craft's progress. The vortices of a fishtail, in contrast, rotate in a direction opposite to those from an object such as a paddle. The Triantafyllou brothers, aided by Mark A. Grosenbaugh of the Woods Hole Oceanographic Institution, found that the reversal in direction of these tiny whirlpools produces not the drag of the paddle but a jetlike thrust, which accounts for the speed and rapid acceleration of many fish and cetaceans. These experiments were done with a mechanical device that reproduces the taillike movement of a fish. The data will be refined with the complete robotic tuna.

A figure called a Strouhal number characterizes the fluid dynamics of the turbulent wake. The optimal number to give a fish a boost was about 0.3. The researchers inspected the literature for more than 15 species of fish and cetaceans, from dace to dolphins. All except the saithe had a Strouhal number from 0.25 to 0.35. "We've shown that you get your best buy for your money at these numbers," George Triantafyllou says. "Fish have been swimming for millions of years. They have

had plenty of time to discover this."

The tail movement also converts a hindrance to an advantage. When a fish moves through water, its undulating body produces drag. Positioning the tail at a certain angle with respect to the body reduces the resistance created by drag-inducing vortices, thereby producing thrust. The workers have also filed for a patent on a fishtaillike control surface that flaps, enabling a submarine or even an airplane to climb or descend at high angles of attack without lapsing into an uncontrollable stall.

RoboTuna may also help the researchers discover why some fish can veer sideways within a radius of 10 percent of their length. The most maneuverable ship takes three or four lengths to make a turn. A taillike movable foil would also produce much less noise than a propeller, an obvious benefit in submarine design.

The Office of Naval Research recently sponsored a workshop at Woods Hole that brought together investigators with an interest in how fish move. It could be a prelude to a navy-backed cross-disciplinary program—hydrodynamicists and fish biologists taking group swimming lessons from a tuna or jack mackerel.

—Gary Stix

SCIENTIFIC AMERICAN

**COMING
IN THE
FEBRUARY
ISSUE...**

SULFATE AEROSOL AND CLIMATE

R. J. Charlson
*University of Washington,
Seattle*

T. M. L. Wigley
*University Corporation for
Atmospheric Research,
Boulder, Colorado*

GLOBAL STRATEGIC ANALYSIS

**Philip Morrison,
Kosta Tsipis and
Jerome Wiesner**
*Massachusetts Institute
of Technology*

LIQUID MIRRORS

Ermanno F. Borra
Laval University

ALSO IN FEBRUARY...

Homeotic Proteins and
Development Design

Science in Pictures:
The Reconfigured Eye

The International Problem
of HIV and AIDS Among
Injecting Drug Users

The Terror Birds
of South America

Trends: Theoretical
Metaphysics

**AT YOUR
NEWSSTAND
JANUARY 25**

Survival Tactics

*Japanese research managers
huddle closer to the market*

Bouts of economic malaise and corporate restructuring have caused U.S. management to rethink the role of the corporate research laboratory—mostly to the detriment of those institutions and research funding in general. Have Japanese managers, now battered by two years of recession, found different strategies for adjusting research and development programs to economic adversity?

Few big Japanese corporations have drastically curtailed research efforts. The link between technology and the country's dramatic postwar success is too well recognized for that. Instead some Japanese high-technology firms are spending somewhat less, and many are reallocating resources from basic research to product improvement and manufacturing. As Mitsubishi spokeswoman Alison Clark puts it: "It's the need for profit soon. We're taking a more practical view now."

Mitsubishi spent about U.S.\$1.59 billion (172 billion yen) on R&D during the 1993 fiscal year, which ends March 31, and will spend about the same amount this year. (Inflation, which is low in Japan, will exact only a small real decrease in the outlay.) But expenditures will skew toward new product development, primarily in areas such as bringing 64-megabit DRAMs into production and creating advanced office computers and car navigation systems.

Hitachi had outlays of \$3.66 billion on R&D in fiscal 1992. In 1993, \$3.45 billion was budgeted—nearly a 6 percent decrease. Major parts of the research budget will support development of advanced memory devices and computers. A market-oriented system called Strategic Business Projects established a year ago governs Hitachi research activities. The approach is an attempt to speed products more rapidly to market. Toshiba will also lay out slightly less on R&D this year—\$2.50 billion in fiscal 1993 as opposed to \$2.55 billion in 1992.

Fujitsu, however, will sharply cut its research budget this year—by \$410 million. According to corporate spokeswoman Yuri Momomoto, research expenditures have been trimmed to \$2.54 billion. "We are reviewing R&D items now and trying to be more selective and more product oriented," she explains. In spite of the cutbacks, Fujitsu is still spending more than 10 percent of net sales on R&D. Last year the corporation established a system to encourage creativity in research. The "My Way Project" allows researchers to create and propose their own research theme. If approved by a screening committee, the researcher will be given the time and money to work on the project. Three years is customary, but even if there is no final result, the researcher will be granted more time if the project appears promising.

The founders of Sony Corporation built the firm on research. Yet the management planned to increase research expenditures only slightly, to \$2.17 billion, compared with \$2.15 billion in

Japanese Electronics Slump

Forecast for fiscal 1993 (through March 31),
in millions of dollars, with accompanying annual percentage change

	R&D SPENDING	SALES	PRETAX PROFITS
FUJITSU	2,545 -13.8%	19,991 -9.9%	250 *
HITACHI	3,453 -5.7%	34,244 -3.0%	555 -23.1%
JVC	120 -16.7%	4,674 -1.6%	-231 *
MITSUBISHI ELECTRIC	1,591 1.1%	22,212 -3.8%	259 -13.4%
MATSUSHITA ELECTRIC	3,795 1.9%	40,259 -4.4%	583 -34.9%
NEC	2,591 -3.4%	26,840 1.1%	278 65.7%
SHARP	1,018 6.2%	10,643 -0.2%	380 -7.9%
SONY	2,175 1%	16,289 -5.8%	296 -29.9%
TOSHIBA	2,499 -1.9%	30,079 3.2%	417 -17.9%

SOURCE: Nikkei Weekly and company sources

*Fujitsu and JVC had pretax losses in fiscal 1992.

fiscal 1992. Sony carries on research at three levels. A facility in Yokohama houses physicists, chemists and mathematicians. Business groups conduct any research mandated by immediate market conditions. And Sony's Corporate Research Laboratory bridges the gap between basic research and product development, by creating such technologies as magneto-optical systems.

Over the past fiscal year Matsushita Group companies collectively spent \$3.72 billion. For the current fiscal year, the forecast is about \$3.79 billion. Matsushita's Central Research Laboratory in Osaka does most of the corporation's basic research. Divisions are responsible for product development. But company spokesman Akihiro Tanii emphasizes that "Matsushita's aim is not to produce Nobel Prize winners."

And that may be a fundamental flaw in the strategic vision of many Japanese corporate behemoths (one shared by their rivals on other continents). Mitsubishi, Toshiba, Hitachi and other giants make everything from refrigerators to semiconductors. This attachment to a broad array of established products is coming under fire even from the Japanese themselves. One critic is Keiske Yawata, who passed decades with Japan's giant NEC Corporation before leaving to become president and eventually chairman of LSI Logic K.K., the Japanese subsidiary of an American semiconductor firm that specializes in developing ASICs (application-specific integrated circuits). He points to the glut of commodity consumer electronics products. "Everyone already has a VCR," he remarks, "and they never wear out." Yawata feels strongly that Japanese research must become more innovative and more focused to succeed in the future.

For example, Japanese firms chose to put their capital and effort behind the development of memory chips, an established technology blessed with a ready market. Japan became the world leader in memory chips, but now the devices have become a commodity product. Profit margins are paper-thin, and Korean firms such as Samsung provide stiff competition.

Arlon Martin of AT&T Microelectronics Japan sees strong resemblances between the U.S. and Japan with respect to how firms do research. "NEC, Fujitsu and Toshiba have central research labs and also do product development on the local level," he observes. "Projects move from the central lab to a level more dedicated to product design."

But in the U.S., Martin notes, research often provides the opportunity for a business start-up. "In the U.S., a lot of important work is done by start-ups

"The family suggests that memorial contributions be made to the American Heart Association."

When people want to honor a loved one and fight heart disease.

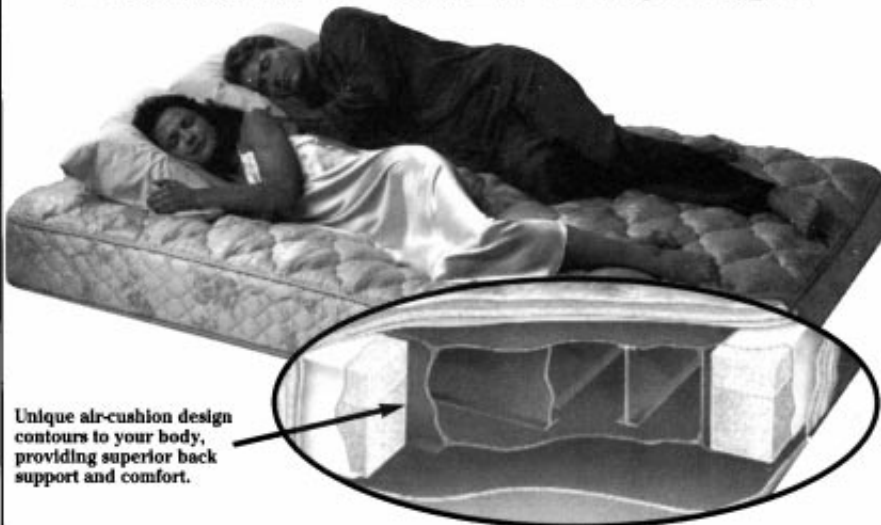
THE AMERICAN HEART ASSOCIATION MEMORIAL PROGRAM®



This space provided as a public service.

Breakthrough mattress technology gives you

Air Support For Back Pain Relief



Unique air-cushion design contours to your body, providing superior back support and comfort.

Only Select Comfort's unique air cushion design allows you to select different firmness levels for each side of the bed.

- Sleep Risk-Free for 90 Nights
- 15 Year Limited Warranty



Select Comfort provides proper back support and contours to your body, distributing weight evenly.



Innerspring mattresses create pressure points and uneven support.



Waterbeds rely on displacement and can cause a "hammock effect," bending your spine unnaturally.

Free Video And Brochure

Call **1-800-831-1211**

YES! Please rush me all the facts on how I can receive a more comfortable night's sleep with Select Comfort's breakthrough mattress technology...and details on your RISK-FREE 90 Night Trial!

Name _____
 Address _____
 City _____ State _____ Zip _____
 Evening Phone (____) _____

SELECT COMFORT CORPORATION

9301 75th Ave. N., Minneapolis, MN 55428-1003
 © Select Comfort Corp. 1994 Dept 2186

and by people coming out of universities, not just in Silicon Valley but in small companies in other areas such as Boston. That kind of research doesn't happen in Japan." But, Martin says, entrepreneurial start-ups do happen in Taiwan: "A bunch of college grads get together and start a new company in a garage. They develop a little expertise, and they're in business."

In Japan, it is much more difficult for a new company to gain the funding and build the key relationships that are so critical to success. Tokyo University graduates simply do not have the same options as do their counterparts in either Taiwan or the U.S. All they can do is choose between companies such as NEC, Matsushita and Toshiba. In the U.S., a thriving high-tech company with innovative technology may be only five or 10 years old. The Japanese business landscape is virtually devoid of such enterprises. The venture capital does not seem to be available.

Not everyone agrees that Japan's basic research performance is deficient. Mark Pearce, a London native who came to Japan to study at a Japanese university, has been with NEC since his graduation five years ago. His is a view of Japan through Western eyes that have learned to see beneath the surface of Japanese society and business. Pearce rejects the view that Japan is far behind the U.S. in basic research. "I think that's changed in the past five years," Pearce asserts. "There's a lot of groundbreaking work now being done in Japan."

Some of the areas in which NEC researchers are working are production of carbon nanotubes, studies of the structure of the nervous system in nematodes, computational chemistry using supercomputers to visualize the movements of single atoms, and melt science—looking at the effect of different gases on the growing of silicon ingots. In Princeton, N.J., NEC is studying language acquisition in children with a view toward the eventual creation of systems for automatic translation.

At the International Solid State Circuits Conference in February 1994, NEC researchers will present seven papers, including one announcing the development of a 64-megabit flash memory device. Moreover, in the past two years, NEC has had 15 papers published in *Nature*, more than any other Japanese industrial company and only slightly less than the 18 published by prestigious Tokyo University.

Such commitment to research may provoke a few tears across the world in Murray Hill. Time will tell whether it can survive the long Japanese economic winter. —Robert Patton, Tokyo

Material Advantage

IBM pushes silicon-germanium chips into the marketplace

Electronics manufacturers have never fallen in love with gallium arsenide, a material that furnishes faster speeds than the industry's old staple, silicon. Even inveterate speed junkies such as chip makers are loath to give up billion-dollar investments in silicon semiconductor manufacturing facilities for a wholly alien fabrication process. The arrival in the marketplace of an alloy of silicon and germanium means they may not have to.

Workers at IBM and Analog Devices report that a collaboration between them has produced the first soon-to-be commercial device made from such an alloy. They were scheduled to provide details of their work at the International Electron Devices Meeting, held in early December. "This is a real technology road map to drive silicon for the next five, 10 or 15 years," says Bernard S. Meyerson, a fellow at the IBM Thomas J. Watson Research Center and leader of the team that developed the manufacturing technique.

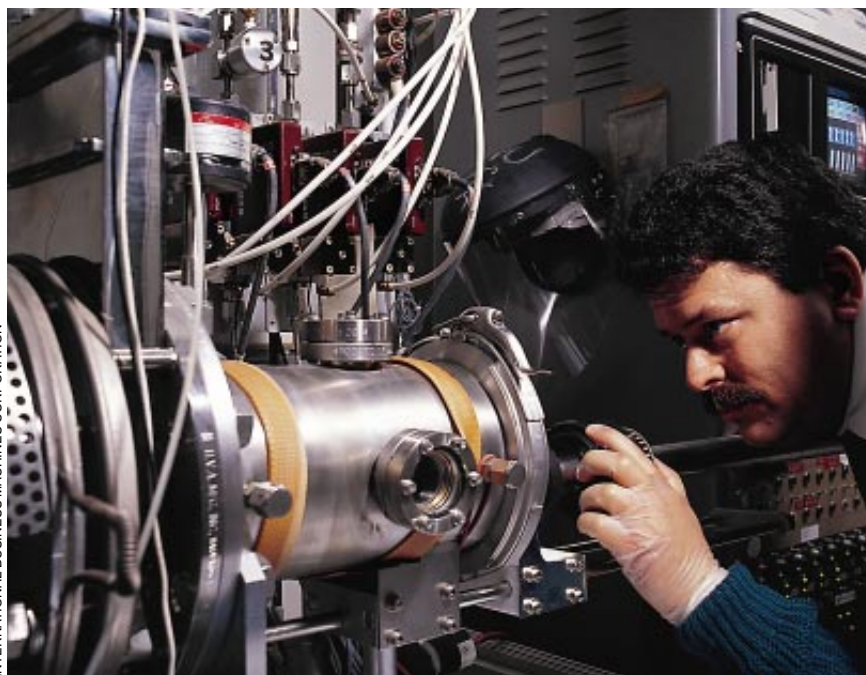
Meyerson and his colleagues describe a bipolar transistor whose base consists of the silicon-germanium alloy. The base turns such miniaturized switches on and off. The purpose of the device is to convert a digital signal into

its analog form at a rapid-fire one billion cycles (one gigahertz) per second. The 12-bit digital-to-analog converter operates faster than any silicon converter to date. It uses less power and is comparable in speed to the best signal converters that use gallium arsenide.

During 1994, Analog Devices, a supplier of analog and digital signal processing semiconductors, will begin selling these building blocks for the digital networks that telephone and cable companies intend to build to every home and business. A signal converter could take a fiber-optic digital signal transmitted to the home and translate it into the analog reception required for a television set. (IBM makes the chips; Analog Devices designed the converter.)

Unlike gallium arsenide, silicon-germanium devices can be made by the same processes as the most widely used integrated circuits are. Because of chip fabricators' long experience with silicon, electronics from silicon-germanium may attain higher speeds at the same cost as other conventional bipolar silicon circuits. "Millions of Ph.D.-hours have been devoted to silicon, in contrast to a tiny fraction of that number for gallium arsenide," says Charles K. Fadel, corporate marketing manager for Analog Devices. To date, NEC, Daimler-Benz and other electronics giants exploring the technology have made individual transistors but have yet to report on integrated circuits, Meyerson says.

Each IBM-made silicon-germanium



BERNARD S. MEYERSON, an IBM researcher, inspects an ultrahigh-vacuum chemical vapor deposition machine, the technology he invented to deposit atomic layers of a silicon-germanium alloy onto a silicon chip.

converter contains 3,000 transistors and 2,000 other elements, such as resistors and capacitors. That number is small compared with the more than one million transistors that a digital chip may sport. But for analog circuits, it marks a high level of circuit integration.

IBM has tried to take advantage of the fact that the high electric field from the alloy causes electrons to move more rapidly across the base. The electric field results from gradually increasing the amount of germanium across the width of the base. That creates a difference in electrical potential, a kind of slope down which the electrons race.

To wed the silicon-germanium to a silicon chip substrate, Meyerson had to solve a serious problem: the silicon-germanium lattice does not match that of the pure silicon substrate. He coped with the challenge by devising a variation of the common technique known as chemical vapor deposition (CVD). A mixture of gases leaves a thin film of silicon and germanium atoms on top of a silicon substrate. An ultrahigh vacuum enables the process to take place at 500 degrees Celsius, less than half the temperature at which a conventional silicon CVD system operates. In this relatively mild environment, researchers can control the precise buildup of each layer of atoms. The silicon and germanium atoms tend not to stay put at the higher temperatures normally used in CVD.

The future looks bright for germanium growers. IBM reported last summer that the maximum frequencies attained by an individual silicon-germanium transistor range from 110 to 117 gigahertz. Such frequencies are more than two times higher than those attainable by a silicon transistor, and they compare favorably with those of many gallium arsenide chips. Analog Devices now contemplates using the silicon-germanium alloy in the electronics for a digital cordless telephone that operates at frequencies as high as three gigahertz (the rated speed of circuits using transistors is always much less than the free-spooling top figure for an individual transistor). The circuitry for sending and receiving radio signals might reside on a single chip, instead of consisting of an amalgam of discrete transistors and capacitors.

IBM researchers have also reported that the silicon-germanium bipolar transistors can be integrated into a chip bearing complementary metal-oxide semiconductor (CMOS) technology, the inexpensive logic and memory devices that are the mainstay for most electronics in personal computers and communications equipment. The CMOS components would serve as the micropro-

cessor and memory; the bipolar circuits would function as high-speed processors of radio signals that may travel, say, from a portable telephone to a communications base station.

The silicon-germanium project marks

a new era at IBM. "In the past, we would have sat on this technology until 1998," Meyerson says. That quickness may help keep alive material designers' love affair with one of the geosphere's most abundant elements. —Gary Stix

Optical Tomography

Light begins to shine as a noninvasive imaging tool

As a child, you may have placed your fingers or palm over a flashlight and looked with fascination at the translucent red glow under your skin. Today researchers are attempting to exploit this simple phenomenon. They seek to create optical imaging devices that rely on the way living tissue absorbs, deflects or scatters red and near-infrared light to reveal structure, density and even physiological processes.

This novel technology, called optical tomography, makes use of the fact that all features of living tissue, from cell organelles to oxygenation levels, transmit, scatter, absorb or otherwise affect photons. So, by measuring the intensity and other characteristics of such light and feeding the information into a computer, images can be composed of what sent the photons flying the way they did.

For now, such optical instruments

cannot hold a candle to the resolution of other methods. Positron emission tomography (PET), magnetic resonance imaging (MRI) and computed tomography (CT) all provide clinicians and scholars with an unprecedented view of the interior of the human body, at very low risk to the patient. Yet these types of scanners require that the patient be brought to a large apparatus.

Optical tomography, on the other hand, may offer a convenient way to monitor tissue function continuously at the bedside, says David A. Benaron, a professor of pediatrics and neonatology at the Stanford University School of Medicine and Stanford's Hansen Experimental Physics Laboratory. Moreover, Benaron and other enthusiasts assert, these devices could produce images faster and less expensively than do the other nonsurgical options.

Some of the enthusiasm may be justified. Optical imaging technology has been showing some precocity. Several teams have already employed optical methods to detect physiological changes in the brain related to hemorrhage, hypoxia and cognitive function. But the



PICTURE OF A CAT, sketched 1.2 centimeters wide on a cellophane sheet, is hidden in an opaque liquid, nine centimeters thick. The cat is displayed on a computer screen using an optical imaging device at the C.U.N.Y. Center for Advanced Technology, directed by Robert R. Alfano.

brain is not the only organ that scientists hope to examine using optical imaging tools. They are developing instruments that would take nonsurgical "optical biopsies" throughout the body. "This is all based on our understanding of what light really does when it is buzzing around in tissue," says Britton Chance, a pioneer in optical technology at the University of Pennsylvania.

To succeed, these devices must somehow contend with measuring the chaotic path of photons scattering through the human body. Most photons scatter hundreds if not thousands of times before they exit the tissue and can be detected. To address this problem, Chance and Arjun G. Yodh, a physicist at the university, expose tissue to pulses of oscillating light whose amplitude has been modulated. These pulses together generate a wavelike "disturbance of brightness." If a single pulse encounters an area of relatively high scattering or absorbency, its intensity will be reduced. This change shifts both the intensity and the phase of the emerging pulses with respect to the entering pulses. Such distortions reveal the extent, density and location of the structures that caused them.

Robert R. Alfano, director of the City University of New York's Institute for

Ultrafast Spectroscopy and Lasers, takes a complementary approach. He tries to determine areas that scatter or absorb light by analyzing only ballistic and snake light photons, the earliest photons to emerge. Alfano's group was the first to exploit the information available from this light, which scatters very little. Alfano uses an ultrafast video camera, backed by recording and analyzing equipment that captures only those photons that emerge in the first one to 10 picoseconds after the pulse is emitted. They can then rotate the object with respect to their instrument and create three-dimensional images. By this method, Alfano's group has deciphered the image of structures as small as two millimeters in diameter through a model medium, similar to human tissue, that is seven centimeters thick. The process takes less than a minute.

Benaron and his colleagues developed the second optical imaging device to be clinically tested. Called the time-of-flight and absorbance (TOFA) scope, the machine measures the time it takes for a fixed percentage of the fastest photons to travel through tissue. This technique still avoids the most distracting photons, those that scatter sharply and are the last to emerge. It does, however, include moderately deflected particles.

"The resolution is not nearly as good as conventional x-rays, but it's a start," Benaron says.

Officials at the National Institutes of Health think optical tomography is promising enough to merit some funding. One of these NIH-supported projects seeks to apply optical tomography to the diagnosis of breast cancer in women younger than 40 years. "An x-ray mammogram doesn't distinguish cancers well in young women," Chance points out. The radiation from a traditional x-ray mammogram scatters too much in young tissue to produce an informative picture, he explains.

NIH grants also support Benaron's group, which will soon test in humans a device designed to monitor oxygen levels in the brain tissue of infants. The instrument, woven of fiber-optic threads, would fit comfortably around a baby's head like a sweatband. Evenly spaced sensors in this optical array would pick up scattered signals from light sent along two axes extending from one side of the skull to the other. The coordinates of deoxygenated areas would be found by calculating the path length and the intensity of light passing through the skull in different directions.

Benaron predicts that simple, spectroscopic devices based on this tech-

If not for you,
then for th
people who depen
on you.

nology could be available clinically next year if endorsement from the Food and Drug Administration is forthcoming. Mediscience Technology, a company that owns patents on several of Alfan's designs, is working toward FDA approval. Early versions would probably perform such tasks as monitoring glucose levels in diabetics, Benaron notes.

Yet this technology could be available sooner for application in areas other than clinical medicine, says Eva Sevick, a professor of chemical engineering at Vanderbilt University. "These optical techniques could improve manufacturing in the chemical industries," she says. She is also studying how optical devices could be used to monitor toxic fumes emitted from smokestacks.

Despite their excitement, researchers in this field worry. "Every new technology has promised to be cheaper," Benaron says. "We don't want this to be the new thing that everyone has to have." To avoid such an outcome, the teams studying different facets of this technology are in close communication. "Some scientists say we're going to have optical imaging in a year. I don't think that's going to be the case," Sevick cautions. "It's not certain it will work, but if it does, the payoff will be considerable." —Kristin Leutwyler

Gene Readers

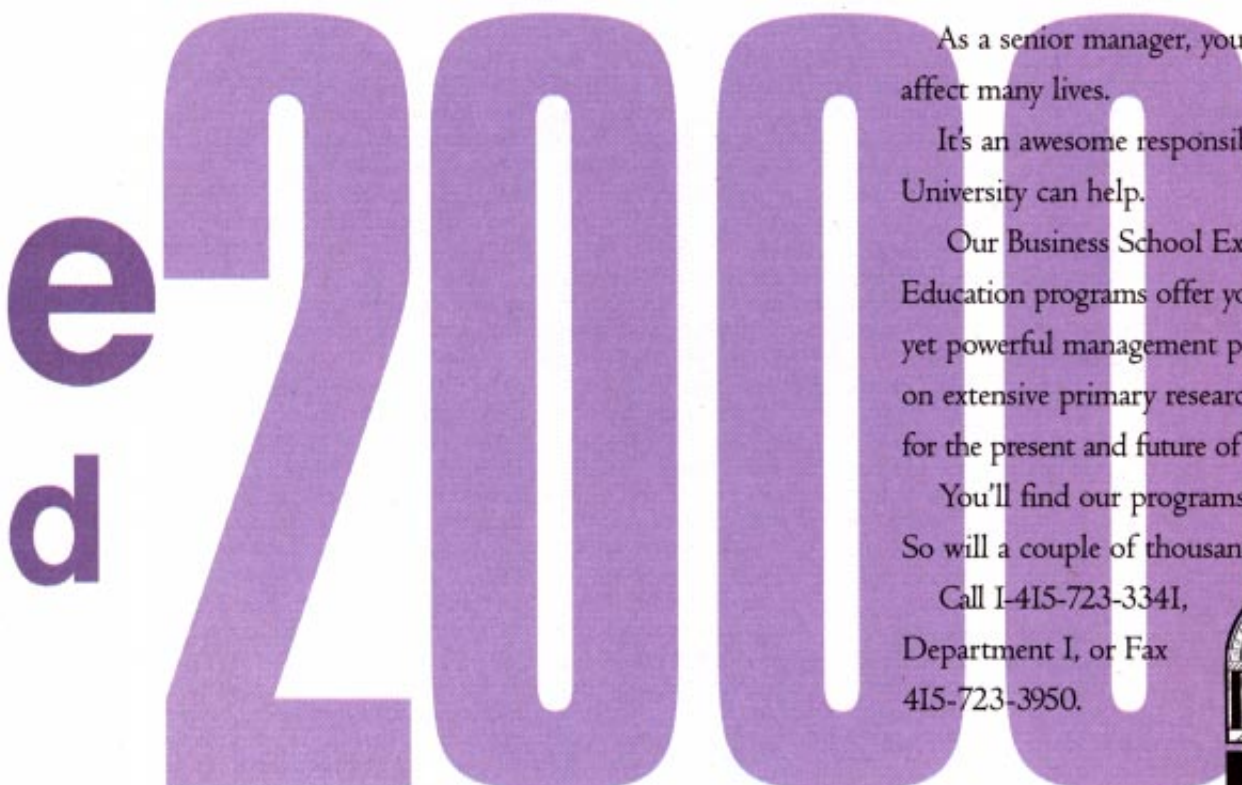
Microelectronics has begun to merge with biotechnology

Two great revolutions of 20th-century science, microelectronics and molecular biology, are joining forces to produce a powerful analytical and manufacturing technology. In the works are centimeter-square-size versions of laboratory instruments that can determine the sequencing of bases in a strand of DNA, analyze proteins or detect human pathogens. "There is an awakening in the scientific community that microlithography, which to date has been the playground of electronics engineers, can be used to make things totally divorced from electronic circuits," says D. Jed Harrison, a professor of chemistry at the University of Alberta in Edmonton.

Harrison's group, working with Ciba-Geigy, claims in an issue of *Science* late last year to have made progress in fashioning an electrophoretic analyzer. Electrophoresis, first conceived in the 19th century, is a primary method for sequencing pairs of bases in nucleic acids as well as for separating the amino acids that make up proteins.

To make the instrument, researchers used an ultraviolet beam to expose a pattern through openings in a stencil-like mask on a glass surface that was to be cut into a centimeter-square chip. Hydrofluoric acid was then applied to etch away the surface to leave a network of channels, or capillaries, 10 to 20 microns deep. To determine the identity of the amino acids, the workers dissolved them in a carbonate solution, which they then introduced into the channels. A positive electrode on one side of the chip and a negative electrode on the other generated an electric field. The field caused the amino acids, which carry small charges on their surface, to migrate through the channels toward the negatively charged electrode. Smaller molecules travel faster, so amino acids of differing density could be separated into discrete bands. Each molecule is tagged with a fluorescent dye.

Harrison's researchers reported that they could separate six amino acids in about 15 seconds, compared with the minutes it takes for existing commercial capillary electrophoresis techniques and the hours for older electrophoresis methods still in use. Harrison hopes eventually to create a general-purpose chemical sensor that can be tailored to ferret out two or three target molecules



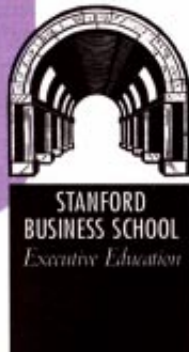
As a senior manager, your decisions affect many lives.

It's an awesome responsibility. Stanford University can help.

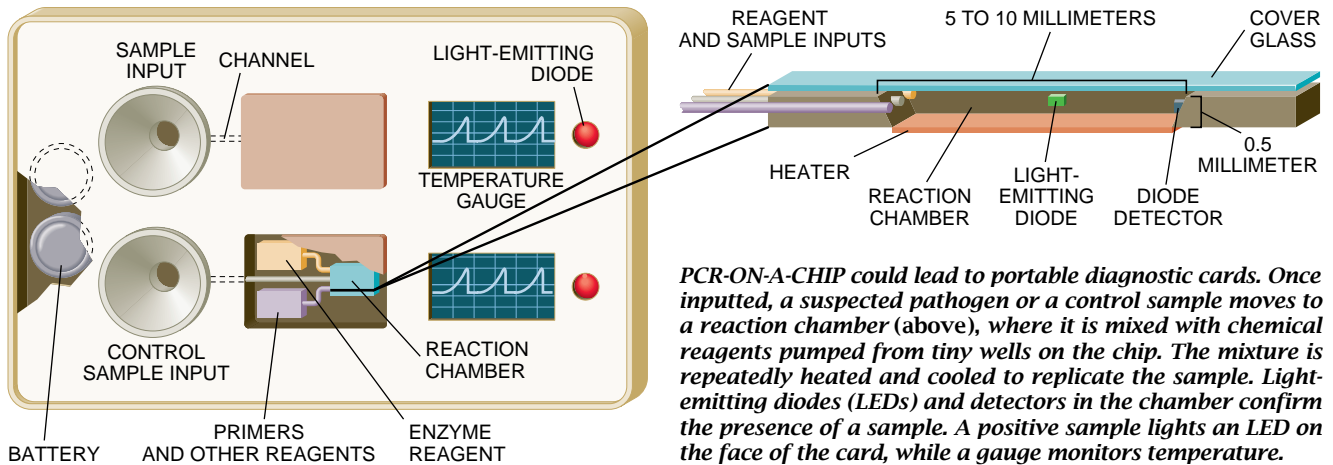
Our Business School Executive Education programs offer you fundamental yet powerful management principles based on extensive primary research. Blueprints for the present and future of your company.

You'll find our programs beneficial. So will a couple of thousand others.

Call 1-415-723-3341, Department I, or Fax 415-723-3950.



STANFORD BUSINESS SCHOOL
Executive Education



from a complex sample. Such a miniaturized instrument will require integration of a power source to generate the high electric fields, light-emitting diodes for energizing the dye, and other diodes to detect the photochemicals.

Minute fluid systems create problems analogous to those faced by engineers trying to connect wires to tiny megabit circuit elements. "Things are done now with macroscopic plumbing," remarks Daniel J. Ehrlich, a micromechanic at the Massachusetts Institute of Technology. "Connectors between tubes are tightened with wrenches. They're all discrete devices like discrete transistors." Ehrlich's project will receive \$762,000 from the Advanced Research Projects Agency (ARPA) over the next three years to build an electrophoretic chip. His work, done in collaboration with the Whitehead Institute, a leading group in the Human Genome Project, will, in part, address moving liquids from macro- to micron-size flow channels. Ehrlich compares it to progressing from a fire hose down to a straw.

Ehrlich's specialty is laser etching. For building prototypes, the laser can make small instruments more rapidly than can photolithography. A laser can etch structures directly into the chip surface. It can also carve in three dimensions and is not limited to the microscopic scale. For example, a millimeter-size well might act as a reservoir for liquids—a structure that would help control the flow between centimeter-size commercial instruments and the microscopic channels located on the surface of an electrophoretic chip.

Such devices, ARPA believes, could also have military uses. The Armed Forces Institute of Pathology wants to identify Vietnam War remains using mitochondrial DNA. A pocket microinstrument could help identify casualties in the field. Lawrence Livermore National Laboratory has received \$1.5 mil-

lion under the same ARPA program for a project to miniaturize the instrumentation for the polymerase chain reaction (PCR), the procedure in which a bacterial enzyme helps to replicate a target sequence of DNA thousands or millions of times. Amplification of strands of DNA would be essential in identifying human remains. For the soldier in the field, a portable PCR kit might yield detectable levels of microorganisms ranging from giardia to the spores of biological weaponry.

Livermore researcher M. Allen Northrup will work with instrument manufacturer Perkin-Elmer and Roche Molecular Systems to build micromachined chambers for heating DNA to separate the molecule into two strands. Once heated, a chamber is then cooled so that stray nucleotides floating around the enzyme-laden mixture build a partial copy by attaching to each separated strand. Before this happens, the segment of each strand to be copied is delineated by attaching a primer, a short length of DNA that marks off the location where the binding of the free nucleotides should begin. As the entire cycle is repeated, the number of copies grows exponentially.

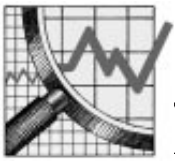
At a technical conference last June, Northrup reported on a prototype for a reaction chamber on a four-inch-diameter silicon wafer. A silicon element on the bottom of the chamber provided heating for the replication of a 142-base pair region of DNA of the AIDS-causing human immunodeficiency virus. Northrup estimates that analysis times may eventually drop from hours to minutes. The size and the low-power requirements presage a credit-card-size portable testing device that could run on lithium watch batteries [see illustration above].

Sequencing the human genome may also lend itself to creative borrowing from electrical engineers. In 1992 the

National Center for Human Genome Research, part of the National Institutes of Health, granted \$2.2-million contracts to Affymetrix in Santa Clara, Calif., and the Houston Advanced Research Center (HARC) in The Woodlands, Tex. (HARC is also one of nine institutions, including Beckman Instruments, that shares \$9.2 million from the National Institute of Standards and Technology's Advanced Technology Program for a Genosensor Consortium to create genetic diagnostic devices with this technology.)

Both HARC and Affymetrix are pursuing separate approaches to sequencing on a chip. Tens of thousands of single nucleotide strands, each eight bases long, will be attached to tiny square areas on a chip surface. Then, samples of DNA, each reduced to a single long strand of nucleotides, will bind in a complementary site on the chip. For the HARC project, electrical or optical detectors in each square will signal that binding has taken place. Deducing the full sequence of hundreds of base pairs on a long strand of target DNA by using eight-base-long nucleotide probes is still a beguiling challenge. Placement of the nucleotide probes onto the chip will also occupy researchers' time.

Some of the instruments under study may ultimately fit together, much as several chips reside on a larger circuit board. PCR-on-a-chip might amplify nucleotides whose identity is then confirmed by microelectrophoresis. Just as with the electronics industry, costs for chemical analysis may drop dramatically. "The cost of automated sequencing machines is typically more than \$100,000," says Lt. Col. Victor Walter Weedn of the Armed Forces Institute of Pathology. "The costs of sequencing for molecular diagnostics may now come down so that the equipment could even be used in a neighborhood clinical laboratory."
—Gary Stix



Lies, Damned Lies and Models

According to the economic models that churned away for more than a year before the House vote, the North American Free Trade Agreement, or NAFTA, is one or more of the following: a brilliant maneuver that will weld the continent into a unified economic superpower, a death knell for the tattered remnants of the U.S. manufacturing sector, a vital boost for U.S. exporters, a roundabout way of redistributing income from the poor to the rich, an economic irrelevance.

Some economists' models predict the disappearance of half a million or more U.S. jobs as a result of lower tariffs on Mexican goods; others predict gains of 100,000 jobs or more, thanks to lower Mexican tariffs on U.S. products. Along with the jobs, billions of dollars in capital may flow south in the coming decades—or then again, they may flow north as Mexico buys U.S. machines for its new factories. Even more perplexing, these same effects might occur as a result of changes in trade policy other than NAFTA.

A computer scientist noting such results would immediately think of the old software adage: "Garbage in, garbage out." Perhaps surprisingly, both the modelers and their critics agree—often cheerfully. The presumptions that have been fed into the models, they say, all too often have little basis in fact.

Among the "little white lies" and "malicious whoppers" (as Paul Krugman of the Massachusetts Institute of Technology calls them) are postulates that employment in each nation will remain precisely constant, that no manufacturing facilities will move from the U.S. to Mexico or (in an opposing model) that every new job south of the border will be counterbalanced by an equivalent one lost north of it. Researchers' figures for the productivity of Mexican workers—a crucial factor in determining just how attractive the country's low wages really are—vary by a factor of three, depending on whether one counts all workers—or just those in new factories built by U.S. firms. And on it goes.

Some observers contend that policymakers should not even bother sorting out the good economics from the bad in this conflicting muddle. "Everybody is just making up numbers," says Jag-

dish Bhagwati of Columbia University [see "The Case for Free Trade," by Jagdish Bhagwati; *SCIENTIFIC AMERICAN*, November 1993].

Indeed, Alan V. Deardorff of the University of Michigan freely admits that many of the methods he and his collaborators used to model NAFTA's effects were "ad hoc." They noticed, for example, that the agreement appeared to increase return on investment in the U.S. and Mexico, and so they assumed capital would naturally flow in from elsewhere to take advantage of higher returns. How much? "We just picked a number" that seemed plausible, Deardorff says.

Gary C. Hufbauer and Jeffrey J. Schott of the Institute for International Economics in Washington, D.C., took a somewhat more fact-based approach: they studied what happened when low-wage countries such as Ireland and Spain joined the European Community

The NAFTA debate was conducted in terms of fallacies exposed 150 years ago.

and extrapolated to project the possible consequences of Mexico's new attachment to the U.S. Although their technique cannot distinguish the effects of NAFTA from those of other developments that stimulate trade across the Rio Grande, Schott says, its overall predictions yield an increase of about 150,000 U.S. jobs over the next decade.

Does that actually mean that unemployment will be a tenth of a percentage point lower in the year 2000? If economists could forecast labor markets that accurately, Krugman scoffs, "the Soviet Union never would have collapsed." Other international events, not to mention the deficit and the Federal Reserve, will swamp any changes that might be attributed to NAFTA.

The message of all the dozens of models is more qualitative than quantitative, Deardorff says. Some predict small positive effects on employment and national income; others predict small negative ones, but all predict

small effects. (Intuitively, this should have been obvious—Mexico's economy and population are small compared with those of the U.S., and tariffs on most products crossing the border are minimal already.) Knowing that the agreement is essentially insignificant as far as the U.S. economy goes, Deardorff says, permits policymakers to concentrate on other issues, such as foreign policy gains or the effects of the agreement on the Mexican economy.

Then again, changes that economists consider small may loom large to voters and lawmakers. As Bhagwati points out, the overall level of employment may hide significant changes in the structure of the U.S. economy—NAFTA is expected to accelerate both the loss of less skilled jobs in the U.S. and the creation of higher-paying, more sophisticated ones. Even if the total number of jobs in the U.S. economy does not change, millions of people will probably be thrown out of work at least for a while. Those in shrinking industries, such as textiles and automobiles, will have to find new jobs in such growth sectors as financial services.

Furthermore, most models suggest that the wages of workers in the remaining low-paying jobs will fall to bring them closer in line with international competition. Mexico will undergo structural changes of even larger scale, as small farmers displaced by competition from Northern agribusiness look for jobs elsewhere.

These dislocations—and the retraining that will be needed for workers to cope with them—should be the primary issues in discussions of NAFTA and its consequences, Bhagwati asserts. Yet they are hardly being debated at all. Why not? His explanation is that U.S. policymakers are simply not prepared to confront them.

Krugman places some blame on economists as well: they have pandered to prevailing fads—"We keep putting 'competitiveness' in article titles" as if it meant something, he comments. And oversimplified arguments have sold well in the public marketplace of ideas, whereas more careful ones sit on the shelf. "NAFTA has made me wonder why we bother trying to learn anything new in international economics when public debate is conducted in terms of fallacies that were disproved 150 years ago," he says.

—Paul Wallich



Knots, Links and Videotape

I've popped the popcorn, Charlotte's brought along the soda—did you manage to rent the video?" Boris asked.

"Of course I did," Alison replied. "*Exterminator 4*."

"Oh, that Ernie Scrambledegger," Charlotte sighed.

Boris put the tape on. The screen showed a pattern of irregular white lines, and then the title sequence came up. They stared at the screen. "Uh—maybe it's an ad," said Alison hopefully.

"I don't think so," Boris said. "Alison, what the devil is *Not Knot*?"

"Whatever it is," said Charlotte in menacing tones, "it certainly isn't *Exterminator 4*. Alison, did you forget your glasses again?"

"Well, there was a bit of a mix-up at the checkout—"

"Hold it, guys," Boris said. "The video store's closed now. We'll just have to make do with whatever it is that Alison got by mistake."

On the screen a mass of multicolored wormlike tubes writhed around

like a creature in its last death throes.

"Aliens," said Charlotte firmly. "Okay, I'll settle for aliens."

"I don't think so," Alison said. "In fact, I'm pretty sure I recognize this. It's a math video produced by the Geometry Center at the University of Minnesota."

"A math video?" said Boris in horror.

"Well, we do live in the age of math communicathionth," Charlotte lisped. "They got a big research grant to set the center up; this is just one of the things they've done. Hey, the graphics are good. Look at those transparent boxes. Alison, what's it all about?"

"Some amazing discoveries about the topology of knots," Alison said.

"Topology?" asked Boris, at the same time as Charlotte asked, "Knots?"

"Um. Imagine tying a knot in a length of tubing and fusing the ends together so that the knot can't escape. The question is, Can you recognize when two such knots are equivalent? That is, can you deform one of them into the other by bending or twisting the surrounding space, without cutting or tearing it?"

"How do you tear space?" Charlotte asked. "Or bend it?"

"Imagine it's filled with some kind of very soft, stretchy, squashy Jell-O, and bend or tear that."

"And this is math? Squashy Jell-O?"

"Not only is the universe stranger than we know," Alison said, "but it is stranger than we can know. Math especially. Be thankful we're only filling space with Jell-O." The screen changed to show three linked rings [see illustration on this page].

"That's a link," Alison said. "Like a knot, but with more tubes. Those are the Borromean rings, which are famous because no two of them are linked, but all three are. I mean, if you cut any one of them, the whole thing falls apart, but if you don't, it hangs together."

"Oh." They watched for a while. "What's this bit about?"

"They're showing you that if you forget about the knotted tubes and just look at the spaces outside them—their complements, that's the word—then inequivalent knots have inequivalent complements."

Charlotte thought for a moment. "Isn't that obvious? If I follow you correctly, the complement is like the whole of space filled with Jell-O but with a tunnel burrowed through it where the knotted tube would be. If you can bend the space outside one knot to look like the space outside the other, doesn't the knot kind of get carried along?"

"Yes, but it might get kind of twisted up, I suppose," Alison said. "Anyway, it can't be obvious, because the same statement is false for links. You can find links whose outsides are the same but whose insides are different [see box on opposite page]."

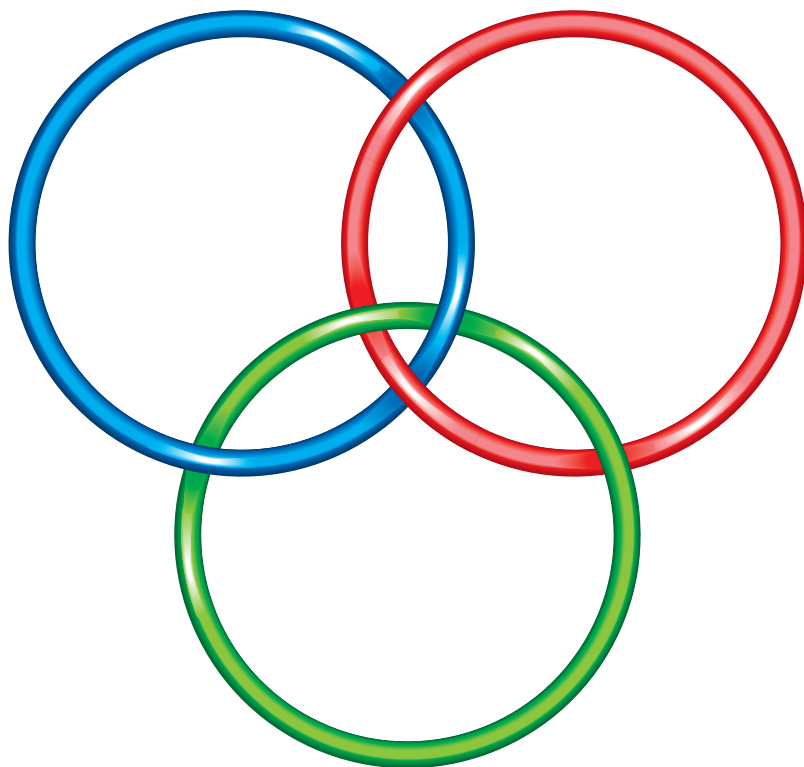
"The argument seems to involve cutting things," Boris protested. "I thought you weren't allowed to cut the space."

"Okay, so I lied. You can cut it, provided you glue it back together later. Just as it was."

"But then you can untie a knot," Boris protested. "Cut the tube, undo the knot, then glue it back."

"Which is why, Boris, dear, you have to deform the space around the knot and not just the knot," said Charlotte, who was finding the video more to her taste than she'd expected. "Which in turn is why it's called *Not Knot*."

"Thanks, Charlotte," Boris said. "Hmm, it's getting more complicated now."



The Borromean rings

“Yes. It’s leading up to the recent discovery that knot complements have a natural geometric structure, which you can use to tell the difference between inequivalent knots. The interesting thing is that it’s non-Euclidean geometry that shows up,” Alison explained.

“You’re telling me there’s more than one kind of geometry?” Charlotte asked.

“There are lots of different kinds of geometry—ordinary Euclidean geometry is just one of them. The main difference in non-Euclidean geometry is that parallel lines behave in funny ways and may not exist at all. You can visualize two-dimensional non-Euclidean geometry by replacing the plane with curved surfaces, like spheres or saddle shapes, and drawing the lines and stuff on those. But for knot complements, you need to think about three-dimensional curved spaces, and that’s hard. So what the video does is fly you around inside such a space and show you what it would look like.”

Boris looked at the screen, where tiny cars were chasing one another around a cone. “Flies? With a car?”

“The flying bit comes later. Here they’re showing you how to make non-Euclidean geometries by kind of cutting a slice out of ordinary space and gluing the edges together. Just as you can make a cone from a circular piece of paper with a pie-shaped slice cut out. Only you just have to imagine what the result of such a gluing process would be. The idea is that whenever you draw a line that hits an edge of the slice, you immediately transfer to the corresponding place on the other edge of the slice and carry on drawing. That kind of bends the lines, even though each bit of them is straight—so not only do you get funny effects with parallel lines but also a ‘straight’ line that can bend around and cross itself.”

Boris looked puzzled. “Alison, how can a straight line bend?”

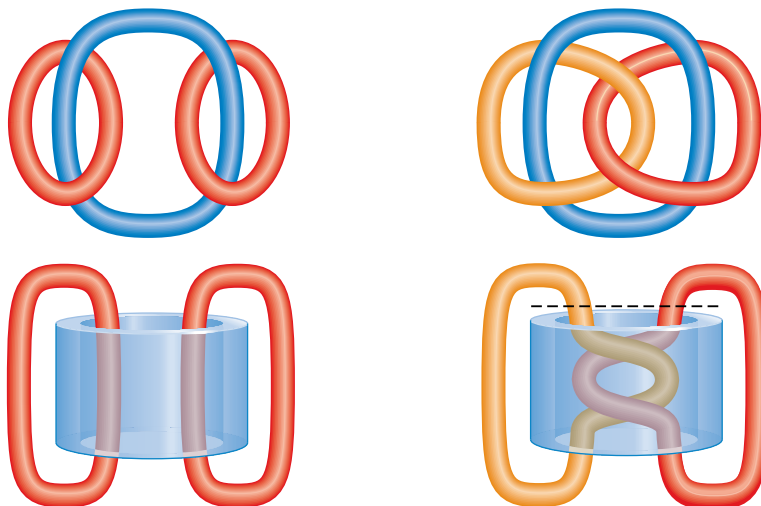
“When I say ‘straight’ I mean the line that covers the shortest distance. When the space has a non-Euclidean geometry, straight lines don’t always look straight to us—looking in from the outside. But if there were light rays that followed the shortest paths, then to a creature living in the space the lines would look straight.”

“Another way to create non-Euclidean geometries is to use mirrors,” Alison went on. “Like a kaleidoscope. The mirrors have the same effect, changing the direction of light rays. For instance, imagine you are inside a cubical room whose floor, walls and ceilings are all mirrors. What do you see?”

Charlotte thought for a moment. “Lots of copies of me.”

Inequivalent Links with Equivalent Complements

The two links at the top are inequivalent. In this topologically equivalent representation (*bottom*), one link has been stretched to create a thick cylindrical tube. Cut the right-hand picture across the top of the cylinder (*dotted line*). If you rotate the links to untwist them and then reattach them, you get the left-hand picture. This proves that the complements are equivalent. But the left-hand link falls apart completely if you cut the cylindrical tube, whereas the right-hand link does not, so the two links are inequivalent.



“Yes. The images of the cube in the mirrors would tile space, and in whatever direction you looked, you’d see yourself. Well, a reflection of yourself. But mathematically we can pretend that each reflection really is you, the same you. That has the effect of ‘gluing’ opposite mirrors together, just as we glued the edges of the cone. But now you get a three-dimensional space with weird geometry. For instance, the straight line starting at your forehead and traveling horizontally forward eventually runs back into your forehead.”

“So straight lines can make U-turns?”

“Yeah. But they stay straight, which is perfectly natural if you happen to be a creature that lives in a knot complement. Now, you can set up a different kind of mirror—mathematical, not physical—that turns things upside down as well as reflecting them. Call them inverting mirrors, okay? If the walls of the cube were inverting mirrors, you’d see lots of copies of yourself, but some of them would be upside down [see top illustration on next page]. This particular geometry, a cube with inverting mirrors for faces, is one possible geometry for the complement of the Borromean rings. The video explains why in detail, but I’ll try to summarize. An inverting mirror—in the sense we’ve just been talking about—is a place where space wraps around itself in strange ways. A kind of space

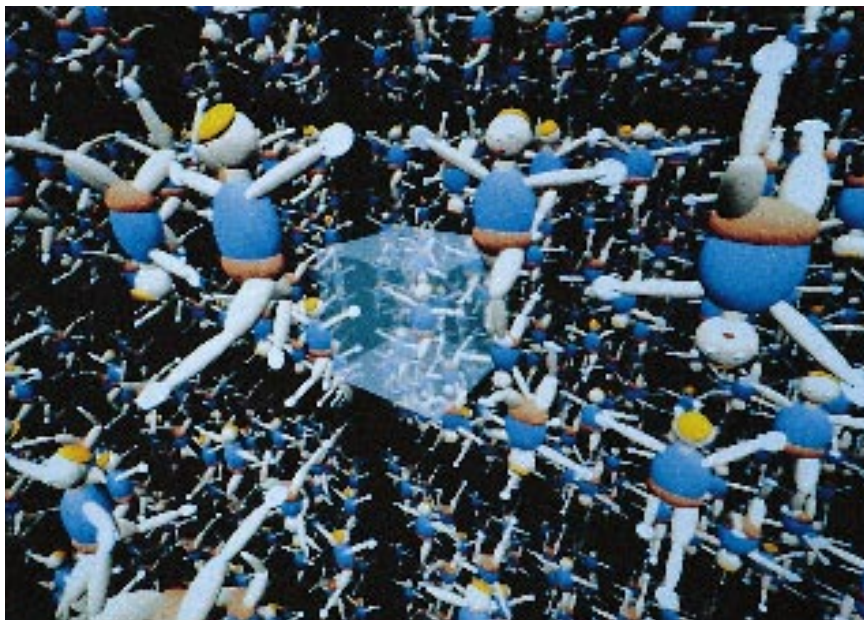
warp. With me so far?”

“Hanging on by my toenails.”

“Now, take the three separate tubes that make up the Borromean rings and stretch each tube until it gets very long and thin, with its edges straight for most of their length. Like an athletics stadium, only with the straight parts of the track being much longer. If you do it right, you can get a pair of parallel tubes that runs north-south, say; another pair that runs east-west; and a third pair that runs up-down. Plus extra U-shaped pieces joining them at each end, which we’ll push away to infinity so they don’t matter.”

“Gotcha.”

“You can find a similar arrangement of lines on a cube. On the floor and ceiling, run a line down the middle, going north-south. On two of the walls, draw horizontal lines running east-west. On the remaining walls, run vertical lines up the middle. The important thing is that those lines stay fixed when you ‘reflect’ them in inverting mirrors on the cube’s faces. That lets you relate the corresponding geometry to the complement of the Borromean rings. It means that when you perform all the inverting reflections, the images of the lines on the cubes fit together so that they stretch away to infinity, just as the Borromean rings do after we push the U-bends off to infinity. And that means that the space around them—the funny



Inverted reflections in all directions: one possible geometry for the complement of the Borromean rings

geometry with lots of copies, some upside down—is just like the space around the Borromean rings.”

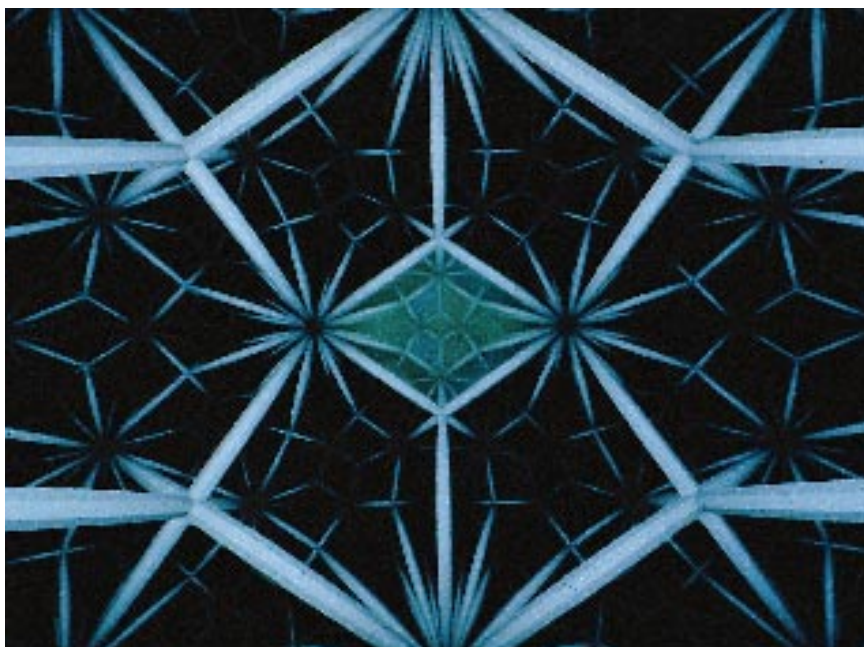
“I’ll believe you,” Charlotte said.

“Okay—or else watch the video again, carefully.”

“The night is young yet,” Boris said. He stopped, spellbound. “Hey, this bit is really neat. It’s like being inside some kind of cage and moving around through the bars.”

“This is the flying part,” Alison explained. “The bars are the edges of dodecahedrons. Well, really they’re all the

same dodecahedron, because you have to imagine everything glued together as the faces of the cube were. But if you lived inside such a space, you’d see multiple copies of everything, like the picture with the inverting mirrors. It’s another geometry you can get from the complement of the Borromean rings by using a different kind of ‘mirror.’ Regular mirrors on the faces of a cube produce cubic images that tile space, and similarly non-Euclidean mirrors on the faces of a dodecahedron produce dodecahedral images that tile non-Euclid-



A non-Euclidean tiling of space by regular dodecahedrons

ean space. The edges are at right angles, so you can fit four of them together, and that’s why they tile.”

“Hold it,” Charlotte interjected. “A dodecahedron I understand—it’s a solid with 12 pentagonal faces.”

“Right.”

“But they don’t meet at right angles.”

“Not in Euclidean space, no. But this space is non-Euclidean—curved, if you wish. Bent just enough to make them into right angles.”

“I guess I see that.”

“The video really does let you see it. You fly through it; you can see what it looks like, really feel the weird curvature effects that come from non-Euclidean geometry [see bottom illustration on this page]. It’s quite strange and beautiful.”

“Yeah,” Boris said. “It kinda gets to you after a while. I could imagine living in a space like this. It’s roomy.”

“How do you mean?”

“Well, flying around it like this you can see that surrounding each dodecahedral tile there are a lot more other tiles than you could fit into ordinary Euclidean space. The amount of space gets bigger than you’d expect as you move outward.”

“That’s called negative curvature. It shows you’re in what’s called hyperbolic space, one particular kind of non-Euclidean geometry. And that’s the central point of the video. According to a recent discovery made by William P. Thurston of the Mathematical Sciences Research Institute, nearly all knot and link complements have a natural hyperbolic geometry. There are a few exceptions, but they’re all known. And you can use the geometric structure to tell all of the others apart. That is, inequivalent knots or links have different geometric structures. It’s an amazing connection—I dare not say ‘link’—between the flexible geometry of topology and the rigid geometry of non-Euclidean spaces. So now a very old-fashioned branch of math—non-Euclidean geometry—is back in vogue.”

“Great,” said Charlotte, rewinding the tape. “But what would be really great is if they hired Ernie Scrambledegger to star in the sequel.”

“The Geometry Center got a big grant,” Alison said. “But not that big.”

FURTHER READING

NOT KNOT. Video. Directed by Charlie Gunn and Delle Maxwell. Supplement written by David Epstein and Charlie Gunn. Produced by Geometry Center, University of Minnesota, 1991. Distributed by A K Peters, Wellesley, Mass. Telephone: (617) 235-2210.



Crossing in Style

LANDMARK AMERICAN BRIDGES, by Eric DeLony. Bullfinch Press, 1993 (\$40).

More than 200 fine large photographs document 95 exceptional examples of U.S. bridges, "American engineering's finest hour." The photographs are in both black-and-white and color, with individual descriptions, a page or two in length, for each. Their intent is not only to image and celebrate these structures but to lead to their preservation through official listing in the Historic American Engineering Record and related registries. About 1,000 landmark bridges of every age and kind are included in the appendix. (There are close to 600,000 bridges in the U.S.; two out of five are classed as unsafe or functionally obsolescent.)

The big list is a happy outcome of close collaboration between the American Society of Civil Engineers and the National Park Service. Browsing through this loving and learned selection, made by the chief of the HAER, is a pleasure and an education for any admirer of "pontists" and a satisfying addition to knowing tourism in America. It will give to bridges a deserving share of the pride and devotion we hold for a Mount Vernon or a Monticello.

The most celebrated bridges are here, from the Golden Gate and the Eads Bridge at St. Louis to Hell Gate, the Kill van Kull and the Verrazano Narrows. Let a few less notable morsels stand as a fresh sample of the whole.

In 1813 builder David Shriver, Jr., took the National Road across the Caselman River in western Maryland on an 80-foot span of masonry arch. It was the longest stone arch in the country. "The bridge's setting...stonework, proportions, and detail are extraordinary." A glance at this colorful autumnal image suggests a harmonious painterly scene from classical China.

The longest covered bridge in the world was built in 1862, not in thrifty New England but on the road that the bonanza Nevada silver and its Cornish miners traveled to and from the Comstock Lode and the port of San Francisco. Constructed at the crossing of the South Fork Yuba River, its shingled siding shows clearly the trace of the grand wood arch sheltered within, a rare distinction fit for the 233-foot span.

In 1882 a great timber company built a neat bridge to carry its logging railroad across the McKenzie River in Oregon. It looks like the frame of an open-work box perched on two masonry walls, one on each bank. The railroad is abandoned. But the Hayden Bridge displays "two milestones" from the high era of American Standard wrought-iron bridges. Its repetitive prefab diagonal tension ties and vertical compression columns follow the Whipple-Murphy design, "the truss of choice" country-wide for intermediate spans. The proprietary "Phoenix columns" were shipped complete from their manufacturer in Pennsylvania; they have a very efficient cross section and came with a kit of cast-iron decorations, even to a nameplate and nifty finials for the portal columns.

The Kinzua Viaduct spans a quiet, well-forested valley among the hills of the northwestern Pennsylvania plateau. Once it carried Erie Railroad track half a mile along unencumbered flat deck, in strong visual contrast with the rounded hills. The load is borne on three dozen pairs of thin, well-braced ironwork

columns up to 300 feet high. This was in its day the highest viaduct in the world. One of the two men who designed the bridge in 1900, glider enthusiast Octave Chanute, was both friend and rival to the brothers Wright while they built their Flyer. Today the viaduct is the focus of a state park; no other piece of human handiwork is to be seen in the entire view, and no noisy locomotive has crossed here for a generation.

Not all the bridges in the book share the strange, lonely beauty of Kinzua. They were chosen rather to display over two centuries the changes of purpose, design, materials and scale, from colonial years to the 1991 act of Congress that authorized flexible measures of preservation. By its design the list does not include noteworthy disasters or the latest structures. A concise timeline of bridge events from Palladio onward is a valuable bonus of the volume.

Conde B. McCullough was state bridge engineer for Oregon during the Depression years. In "a remarkable outpouring of creativity and skill," funded by F.D.R.'s Public Works Administration, he and his staff took scenic Highway



CAST-IRON ARCH BRIDGE in New York City's Central Park, designed by Calvert Vaux and Jacob Wrey Mould, was constructed in 1864.



America's Biggest Selection of Bargain Books

Overstocks, Remainders, Imports and Reprints from all major publishers. Books recently priced at \$20, \$30, \$40—now as low as \$1.95, \$2.95, \$3.95.

Thousands of titles, from yesterday's best sellers to books you never knew existed.

Over 40 subject areas: Science, Biography, History, Fiction, Literature, Sports, Health, Nature and more.

Fast Shipment, normally within 48 hours, and a moneyback guarantee.

Please send me your Free Catalog of Bargain Books.

Name _____

Address _____

City _____

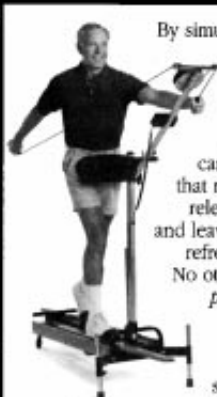
State _____

Zip _____

HAMILTON

Box 15-439, Falls Village, CT 06031

Choose the healthy way to unwind after work.



By simulating the total-body motion of cross-country skiing, a NordicTrack® exerciser provides an invigorating cardiovascular workout that relaxes your muscles, releases pent-up tension and leaves you feeling calm, refreshed and energized. No other exerciser has the patented flywheel and one-way clutch mechanism that provides the smoothest, total-body workout. Best of all, it takes just 20 minutes, three times a week.

30-day in-home trial!

NordicTrack

A CML Company

FREE Video and Brochure

Call **1-800-441-7891** Ext. 320A4

or write: NordicTrack, Dept. 320A4
104 Peavey Road, Chaska, MN 55318-2355

©1994 NordicTrack, Inc., A CML Company • All rights reserved.

101 up the spectacular Oregon coast over 10 major spans. All were built within five years, a family of "some of the...most innovative concrete and steel bridge designs in the world." Seven of them are pictured; one is a pleasing combination of a big through steel arch, its roadway suspended from rods, with a smaller steel deck arch to each side, and two strings of ribbed concrete arches that complete the crossing of Yaquina Bay.

A second interesting bridge on Oregon's 101 is formed by seven open reinforced concrete arches that span the Rogue River. The bridge was the first in the U.S. to use prestressed construction techniques. The arch halves were jacked apart before the concrete was well set, and the gaps at the crown were quickly filled in by concrete plugs. The arch was left in permanent compression, minimizing cracking and allowing a thinner, more elegant, modestly cheaper design, just 20 years after the pioneering Swiss development.

Skywatching

CELESTIAL DELIGHTS: THE BEST ASTRONOMICAL EVENTS THROUGH 2001, by Francis Reddy and Greg Walz-Chojnacki. Celestial Arts, Berkeley, 1992 (paperbound, \$16.95). **THE CAMBRIDGE GUIDE TO ASTRONOMICAL DISCOVERY**, by William Liller. Cambridge University Press, 1992 (\$29.95). **LIGHT AND COLOR IN THE OUTDOORS**, by Marcel Minnaert. Translated and revised from the last Dutch edition of 1968, by Len Seymour. Springer-Verlag, 1993 (\$44.50). **THE NATURE OF LIGHT AND COLOUR IN THE OPEN AIR**, by M. Minnaert. Dover Publications, 1954 (paperbound, \$8.95).

Even if you live beneath the tawny nightglow of some great city, even if you own only the eye lenses you were born with, many *Celestial Delights* can be yours. All you require is a sky not always hidden in cloud. These expert authors assume of you the least of preparations, not even an abundance of free time. If you look up in wonder at the full moon, they can help you to the rest.

The book annotates very simply the rhythm of sun and vagrant moon, through maps of the skyline at notable parts of their unending performance. Then it provides a fine-print summary good through A.D. 2001, giving dates and times of the decisive figures of that dance and of many another. Next, the pages examine the evening and morning stars, once sacred blazing Venus and sunstruck Mercury, and then the

planets beyond, slow to wander, often high in the night sky. The authors attend to a dozen or two bright stars, the gems of selected constellations, wisely conceding that the thousands of naked-eye stars that define the intricate classical sky are too dim now for urban eyes. Our loss of the velvet night is profound, if a little alleviated by a star pattern printed, so to speak, in large type.

In a practical and simple vein, the text, drawings and summaries offer in one deft book a celestial timetable for the rest of the decade (we face a dearth of total solar eclipses, even for the foot-loose). The contemporary cosmos and the directly observable sky of myth are both treated: the Man, the Old Woman and the Rabbit in the Moon; the startlingly repetitive sky track of Venus, so important to Mayan chronologists; a generally good sampling of sky lore from Babylon to the Skidi Pawnee. Meteor showers are not overlooked, and even the less predictable is well introduced: auroras, comets and rare new stars. If this is your year to skywatch, this is your book.

William Liller was for 30 years an accomplished professional astronomer. He retired early to follow his bent, away from tight schedules and knotty measurements, just "to see it first." That prospector's purity has worked well for him; no other name occurs so often in the decade's list of new galactic novae. "I positively guarantee you," says this certified enthusiast of discovery, that a few minutes at dawn or dusk every day with binoculars will reward your patience, some hundreds of hours of it, with a new comet that will forever bear your name.

The fascinating core of this volume looks closely not at the myriad of beginners but at two dozen of the most accomplished of the world's 100,000 telescopists. (A couple of the people here are in fact professional discoverers.) Most remarkable is a quiet parish minister in a small town 50 miles west of Sydney. Amateur Reverend Robert Evans has discovered some 18 supernovae since 1980, against the finding of 30-odd by the world's community of great telescopes and electronic eyes. His mental store of the normal star patterns of a great many galaxies is his strength; he can recognize at once whether a supernova is present. No one else can patrol so many galaxies where supernovae burn; even the computer searches so far are only well-matched rivals.

Carolyn and Eugene Shoemaker, pro searchers, hold the century's record for new comets, 16 of them, the take of seven nights a month at an 18-inch

Schmidt telescope at Mount Palomar. They take about 20 pairs of films nightly, to scan by day, seeking comets and nearby asteroids. She writes: "Too bad there are not more women amateurs doing astronomical things—if they only knew what fun there is in this field!" Liller's text tells much more of the art and the science of search and provides valuable guidance on coordinates, techniques, film, electronics, what to read and where to buy.

Once you are absolutely sure of your discovery, your claim is swiftly circulated to the cognizant world by the Central Bureau for Astronomical Telegrams at the Harvard-Smithsonian Center for Astrophysics, but only after penetrating scrutiny. The bureau receives two or three times as many erroneous supernova reports and four or five times as many comet reports as those that "pan out." The fool's gold of ghost images, emulsion flaws, flyspecks and shifty asteroids must all be washed out, to leave a few glittering flakes of truth. (Brian G. Marsden, appointed assayer, is scrupulously fair, yet he is demanding as to the facts; in his photograph here there is the hint of a smile.)

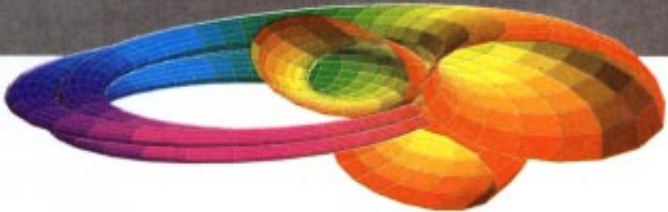
We may witness a celestial event that has no close precedent around July 20, 1994. Several kilometer-size fragments of a newly captured and then tidally disrupted comet should strike the far dark side of great Jupiter about then as they fall from transient orbits. The dusty catastrophe might (or might not) flash out for a few hours as 100-fold brightening of the whole planet! The impact sites will come under direct view hours later as Jupiter spins. No books as yet announce the event, only now crossing the debatable margin of predictability, but prophecies and preparations flower. Watch for more precise news and make ready to turn an eye (some optical glass as well) to Jupiter, come the warm evenings of mid-July.

Astronomer Marcel Minnaert, late of Utrecht, a "spiritual descendent of the Enlightenment," was an authoritative mapper of the solar spectrum, a man at home in 20 languages (even Esperanto), a musician, an outspoken resister of the Nazi occupation and an articulate, reflective lover of nature. His three Flemish volumes on physics in the outdoors are celebrated; the first and best known of them is now out in a splendid new English version.

Minnaert looked hard at the entire visual world and read and thought over what he saw. In some 270 informal sections, mostly on distinct topics, sometimes in more extensive sequence, he heaps up queries, answers, diagrams, challenges and reminders of experi-

Can the most powerful and reliable math software really be the easiest to use?

Macsyma®
A quarter century of software development is hard to beat.
\$349*
Call 1-800-macsyma for a free demo disk.



Engineers and scientists who use Macsyma consistently describe it as more powerful and more reliable than any other mathematics software. Reviewers agree that Macsyma's on-line help system is the best in the field. *IEEE Spectrum* calls Macsyma "a national treasure" and says: "Users with heavy mathematics needs should insist on Macsyma."

And, the most recent PC Macsyma runs fully three times as fast as earlier ones on *PC Magazine's* 1992 benchmark tests.

* PC version in US and Canada. Academic & quantity discounts available. Macsyma is a registered trademark of Macsyma Inc.

Macsyma Inc. 20 Academy Street Arlington MA 02174-6436 / U.S.A.	tel: 617-646-4550 fax: 617-646-3161
	1-800-macsyma 1-800-622-7962

NORDIC FLEX Gold

ADVANCED DESIGN FOR A SUPERIOR WORKOUT

The NordicFlex Gold® strength trainer advances the science of strength training to bring you truly revolutionary results.

GET A FAST WORKOUT, FAST RESULTS.
NordicFlex Gold has a patented isokinetic resistance mechanism that automatically adjusts to the amount of force you apply. According to research, with isokinetic resistance:

- You build lean muscle more effectively than with rubber band systems or free weights.
- You spend less time between exercises than with other systems.

And the ergonomically correct linear exercise motion isolates the muscle groups you want to build for maximum results.

Plus, the NordicFlex® World Class™ Edition offers additional features to enhance your workout, including the electronic performance monitor.



Patented Isokinetic Mechanism



30-day in-home trial
Best of all...it's from NordicTrack!

©1994 NordicTrack, Inc., A CML Company
All rights reserved.

FREE Video and Brochure
Call 1-800-441-7890 Ext. 6K8A4

or write: NordicTrack, Dept. 6K8A4
104 Peavey Road, Chaska, MN 55318-2355

Send me a FREE brochure Also a FREE VHS videotape

Name _____
Street _____
City _____ State _____ Zip _____
Phone () _____

ence, some of unforgettable simplicity, some subtle and elusive. None demands much instrumentation, few calculations beyond simple, artful geometry and estimates of magnitude. Rainbows and their kin and the phenomena of twilight receive extended treatment and now appear in colorful images as well.

When in winter the sun shadows of two bare branches fall together, why does a bright, thin line part the narrower of the two? Why does the window-pane of the bus form a circle of glow around a distant streetlight? How can rocks appear red in a splashing waterfall? Do stars reflect in ponds? Why do we hold a hand above the eyes when we peer into the distance? Seek the famous green flash of the setting sun, the rarer blue and the occasional red one. How can a streetlamp's bright reflection in canal or river waters become misaligned? "Examine systematically the colors of shadows," as the painters do. Connect the real blue sun of 1951 with the colors of the frosted window-pane. Compare any scene viewed both up sun and down (a mirror lets you see both together). Then seek a dozen examples of the difference: they flow from the surface texture of fields of grain, trees in bloom, sea-foam, lakes, snowy roads.

The new edition is spacious, adorned by 50 color plates taken mostly by a remarkable photographer, Pekka Parvainen. Minnaert himself brought jets and day-glow phosphors and Polaroid to update his prewar world of ships and trains and Nicol prisms. But the modest reprint of 1940, plain in grayish black-and-whites, a little dated, a little smaller, remains a bargain well worth a mention. These lines express public gratitude for the unique pleasures of thoughtful observation given so many for so long by M.G.J. Minnaert, born at Bruges just over 100 years ago.

In the Beginning

ORIGINS OF LIFE: THE CENTRAL CONCEPTS, edited by David W. Deamer and Gail R. Fleischacker. With reprints of 46 papers. Jones and Bartlett Publishers, 1994 (paperbound, \$36.25).

Ben Jonson's alchemist held that the gold in the earth was not bred in an instant: "Something came before." This knowing anthology offers an up-to-date view of 20th-century efforts to investigate what came before the living web of earth. The book displays a carefully chosen set of papers, both fragile theory and rigorous experiment, some long familiar,

some unexpected. They are arranged by research themes, along with a concise commentary of some pages for each theme. The work is of value at various levels; not all the intricate geochemistry and biochemistry here is open to general scientific readers, yet a broadly enjoyable book does emerge, as well as a true resource for prepared students. All but a couple of the papers appeared originally in English; the oldest paper cited is from 1908.

The full literature is voluminous; what is selected treats directly one of five themes. They imply a time sequence. First comes the geologic setting four billion years ago, then the prebiotic formation of organic molecules, self-assembling molecular systems, the energetics of such protobiochemistry, and finally the ancestry of the giant informational polymers that now program life. Although the very definition of life employed (no terse definition is really central to so wide ranging a study) specifies the capacity for sustained Darwinian evolution, that epic of the eons is offstage. The genetic code, micropaleontology and the chronicle of repeated unions within and without that led to vital organelles and at last to multicellular organisms are not included, although they are well referenced.

This story centers on ancient single cells or quasi cells, their substance and their actions. Here is a small bouquet of exciting results, many from this past decade, a time of rebirth of interest in and progress toward the grand solution.

Impact! The scarred face of Sister Moon dates the time of massive orbital bombardments. Could life persist while the oceans boiled? A meteoritic autoclave sterilized the earth, relenting only around 3.8 ± 0.2 billion years ago. Whereas soft comet dust and atmospheric shock waves made by large-scale intruders might supply organic monomers, the solar ultraviolet, curiously more intense than though the sun was dim, was a primary source of such material, perhaps even enough to form a dilute organic oceanic soup from the favored prebiotic atmosphere of carbon dioxide and nitrogen. Sea-foam and clay surfaces were present, able to adsorb and concentrate.

One paper appeals by its time-free chemical logic. It seeks to answer why the life we know depends so centrally on the phosphorus atom, oxidized as phosphate. Four properties single out this small radical among its chemical peers. Any linking unit must have at least two valence bonds; the best way to confine such a link within walls of lipid membrane is that it be ionized, calling for a third valence site. But all

other likely multivalent species react too strongly with the watery medium of all life. Thus, the life we know needs the acid radical of P. Phosphorus is, after all, one of the dozen most abundant atoms of the earth's crust. Silicic acid, citric acid, baselike analogues—all fail.

Chemical self-assembly of the right molecules is as plain, as photogenic and as well understood as the familiar formation of froth on soapy water. Dispersal of phospholipids in water spontaneously yields tiny enclosed bags of water within lipid layers. Such bags flatten when dried and naturally contact their neighbors. Rewetting restores larger enclosures, now holding more of the solute. Cycling tide-pool conditions have even shown encapsulation of nucleic acids, a newly direct instance of a process long seen as relevant. It even turns out that chemical energy, the transfer of protons in particular, can arise from the differences in ion concentration between inside and out, and energy can be fed in by photon absorption in pigment within: "a minimum protocell," if hardly an evolving species.

In the end, we come to information storage and its promise of evolutionary reproduction. The novelty of the decade was evidence that a polymeric nucleic acid, a taped program, could also direct chemical change—information and enzyme at once. The molecule that is both chicken-and-egg is not the famed DNA but its more labile kin RNA, at present the translating and transcribing substance of most life. For viruses, it is often the memory tape. But its detailed self-replication seems unlikely. Unwinding the helix is no easy task, and the needed specific monomer choices are hard to make.

If there was an early, simpler RNA world, it was not yet our proteinaceous life. The RNA organisms within lipid walls at subbacterial size had to stretch themselves to extend their catalytic range as the choicest monomers grew fewer. "Once they began to dabble in the use of short peptides," once they found that related DNA supports a far more stable memory, "the RNA world had fallen and the DNA/protein world had risen in its place." So read the last lines of the last reprint, a gripping 1991 paper by Gerald F. Joyce of the Scripps Research Institute.

The time was some 3.7 billion years ago, the place along the shores of the shallow seas. Catalytic RNA did not make itself: the needed helical strings of the right nucleotides seem too long. The gap between chemistry and life has been plausibly shortened, but it is certainly not yet closed. Something came before.



The Tragedy of Enclosure

During the dry seasons in the far northwest of Kenya, the people of the Turkwel River keep themselves alive by feeding their goats on the pods of the acacia trees growing on the river's banks. Every clump of trees is controlled by a committee of elders, who decide who should be allowed to use them and for how long.

Anyone coming into the area who wants to feed his goats on the pods has to negotiate with the elders. Depending on the size of the pod crop, they will allow him in or tell him to move on. If anyone tries to browse his animals without negotiating first, he will be driven off with sticks; if he does it repeatedly, he may be killed. The acacia woods are a common: a resource owned by many families. Like all the commons of the Turkana people, they are controlled with fierce determination.

In the 1960s and 1970s the Turkana were battered by a combination of drought and raiding by enemy tribes. Many people came close to starvation, and the Kenyan government, the United Nations Development Program and the U.N.'s Food and Agriculture Organization decided that something had to be done to help them. The authorities knew nothing of how the Turkana regulated access to their commons. What they saw was a succession of unrelated people moving in, taking as much as they wanted, then moving out again. It looked like a free-for-all, and the experts blamed the lack of regulation for the disappearance of the vegetation. This was, in fact, caused not by people but by drought.

The authorities decided that the only way to stop the people from overusing their resources was to settle them down, get rid of most of their animals and encourage them to farm. On the banks of the Turkwel River they started a series of irrigation schemes, where the ex-nomads could own a patch of land and grow grain. People flocked in. With the first drought the irrigation scheme collapsed. The immigrants reverted to the only certain means of keeping themselves alive in the savannas: herding animals. They spread along the banks and into the acacia woods.

Overwhelmed by their numbers, the elders could do nothing to keep the outsiders away from the trees. The pods and the surrounding grazing land were swiftly exhausted, and people started

to starve. The commons had become a free-for-all. The authorities had achieved exactly what they set out to prevent.

The overriding of commoners' rights has been taking place, often with similarly disastrous consequences, for centuries, all around the world. But in the past two decades it has greatly accelerated. The impetus for much of this change came from a paper published some 25 years ago, whose title has become a catch phrase among developers.

In "The Tragedy of the Commons," the American biologist Garrett Hardin argued that common property will always be destroyed because the gain that individuals make by overexploiting it will outweigh the loss they suffer as a result of its overexploitation. He used the example of a herdsman who keeps his cattle on a common pasture. With every cow the man added to his herds, he would gain more than he lost: he would be one cow richer, and the community as a whole would bear the cost of the extra cow. He suggested that the way to prevent this tragedy was to privatize or nationalize common land.

The paper, published in *Science* in December 1968, had an enormous impact. It neatly encapsulated a prevailing trend of thought and appeared to provide some answers to the growing problem of how to prevent starvation. For authorities such as the World Bank and Western governments, it offered a rational basis for the privatization of land. In Africa, among newly independent governments looking for dramatic change, it encouraged the massive transfer of land from tribal peoples to the state or to individuals.

But Hardin's paper had one critical flaw. He had assumed that individuals can be as selfish as they like in a commons because no one stops them. In reality, traditional commons are closely regulated by the people who live there. Common property has two elements: common and property. A common is the property of a particular community that, like the Turkana of the Turkwel River, decides who is allowed to use it and to what extent.

Hardin's thesis works only where no ownership exists. The oceans, possessed by no one and poorly regulated, are overfished and polluted. Every user tries to get as much out of them as possible, and the cost of their exploitation

is borne by the world as a whole. These are not commons but free-for-alls.

The effects of dismantling the commons to prevent Hardin's presumed tragedy can scarcely be overstated. While their impact has been felt by traditional peoples throughout the less developed world, no group has suffered more than those singled out by his paper: the traditional herders of animals, or pastoralists. In Kenya, the Masai have been cajoled into privatizing their commons: in some parts, every family now owns a small ranch. This has undercut the very basis of their survival.

In the varied and changeable savannas, the only way a herder can survive is by moving. The Masai followed the rain across their lands, leaving an area before its resources were exhausted and returning only when it recovered. Now, confined to a single plot, they have no alternative but to graze it until drought or overuse brings the vegetation to an end. When their herds die, entrepreneurs move in, buy up their lands for a song and either plow them for wheat and barley, exhausting the soil within a few years, or use them as collateral for securing business loans.

Around the world, changes in the ownership of land lie at the heart of our environmental crisis. Traditional rural communities use their commons to supply most of their needs. To keep themselves alive, they have to maintain a diversity of habitats, and within these habitats they need to protect a wide range of species. But when the commons are privatized, they pass into the hands of people whose priority is to make money. The most efficient means of making it is to select the most profitable product and concentrate on producing that. As the land is no longer the sole means of survival but an investment that can be exchanged, the new owners can, if necessary, overexploit it and reinvest elsewhere.

The diverse environments protected by the commoners are replaced with uniform fields of grain or livestock. The displaced people move either to the overloaded cities or into new habitats, becoming poorer as they go, threatening the places they move to, sometimes dispossessing other commoners in turn. For human beings, as for the biosphere, the tragedy of the commons is not the tragedy of their existence but the tragedy of their disappearance.

GEORGE MONBIOT is a Visiting Fellow at Green College, Oxford University.