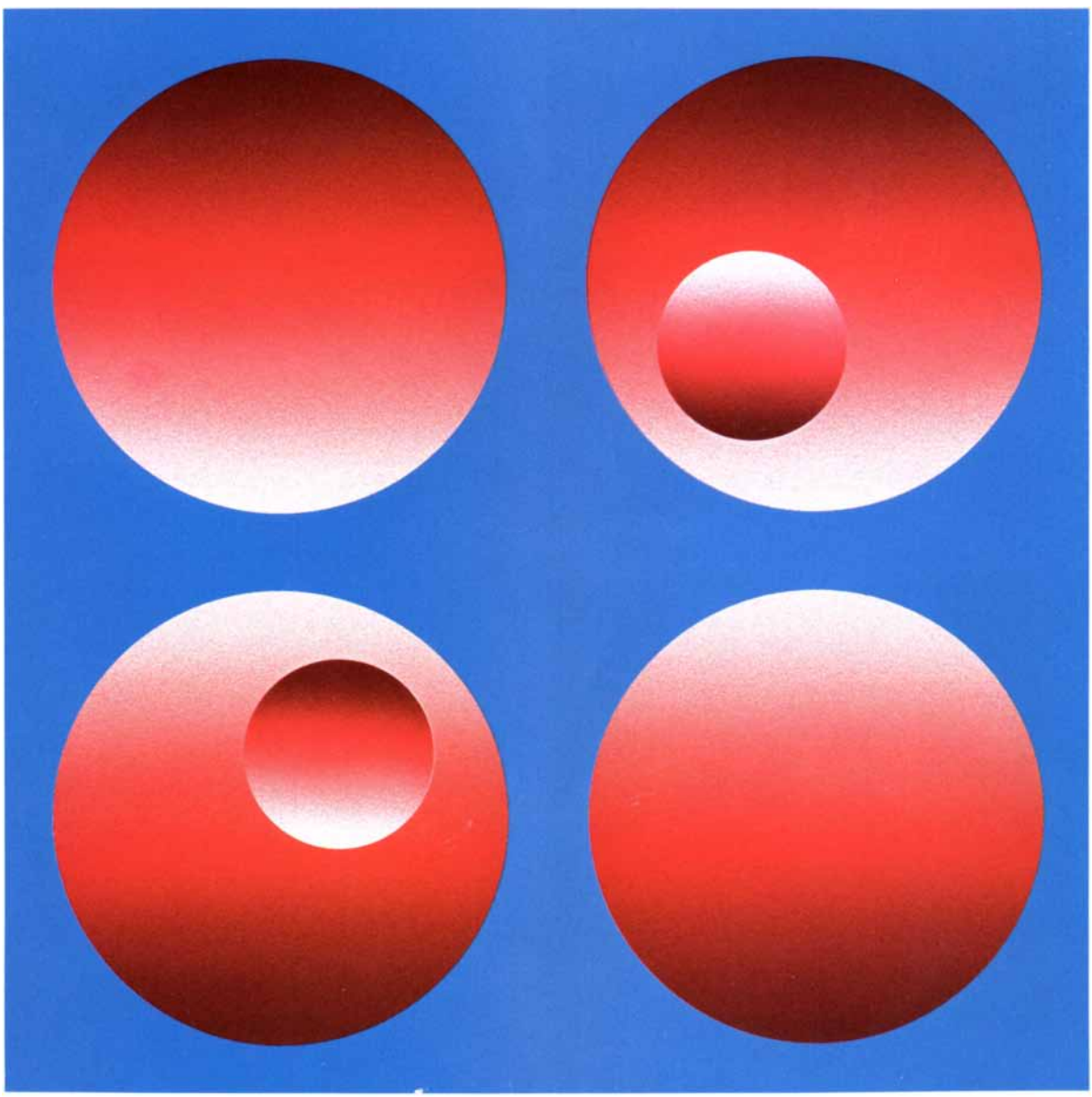
AUGUST 1988
$2.50

SCIENTIFIC AMERICAN

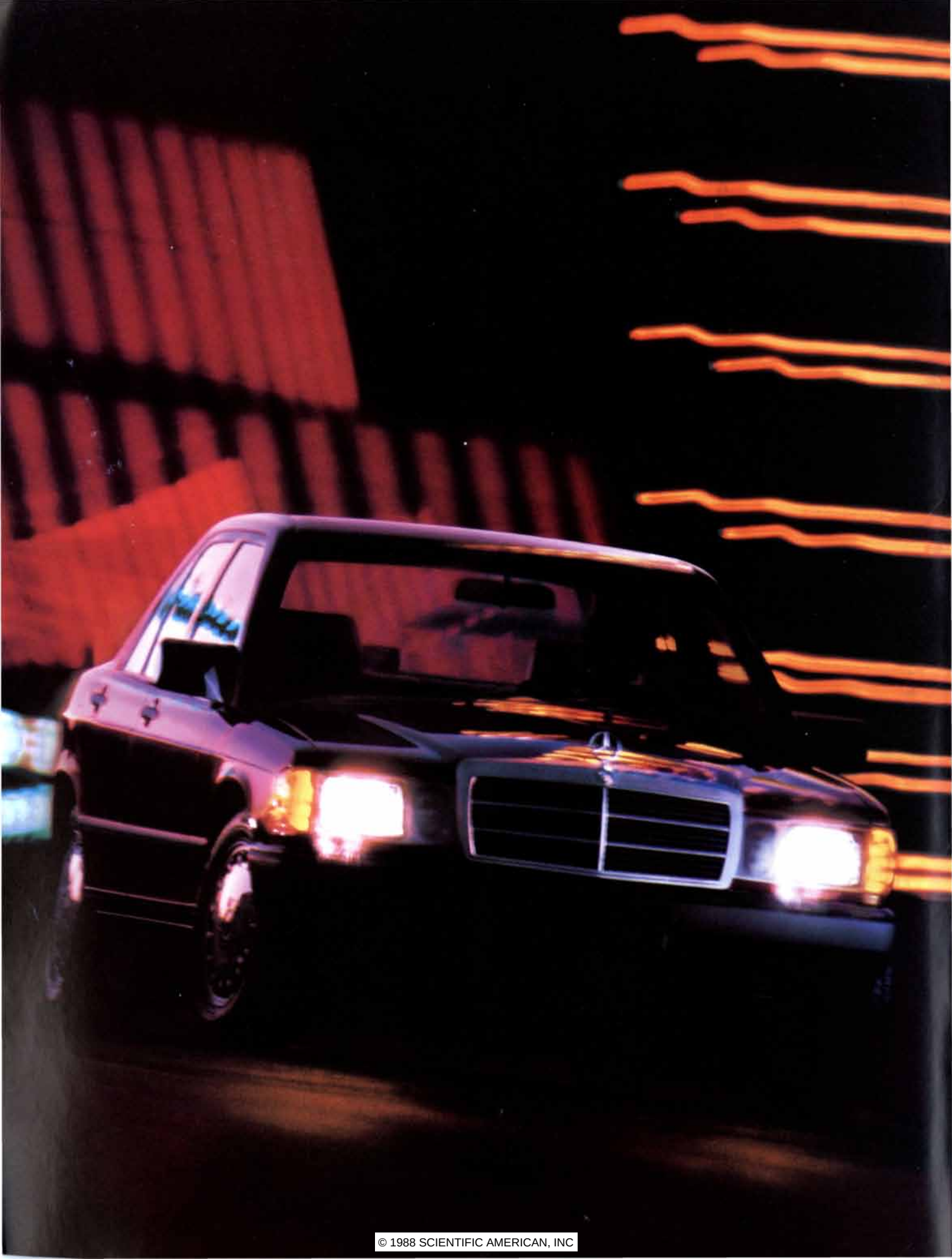
Acid rain's threat—and ways to control it.

How does DNA copy itself with so few errors?

Ultraviolet light activates an anticancer drug.

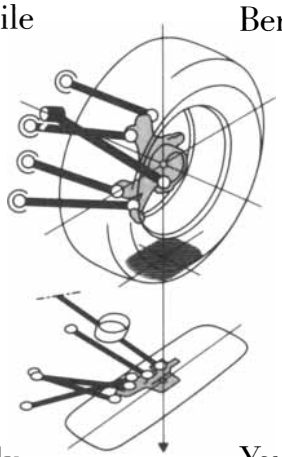


***Perception of depth** from shading depends on the source of the light: invert the page and the hollows become hills.*



THE MERCEDES-BENZ 190 CLASS RESCUES THE SPORTS SEDAN FROM ITS PROLONGED ADOLESCENCE.

Its behavior is a welcome change from the edgy machismo you've come to associate with sports sedans. You feel that this automobile accepts you for who you are—not a frustrated racer but a driver whose enthusiasm for the road is matched by an enthusiasm for the refined.



You wonder, the first time you take the wheel of a 190 Class sedan, how you denied yourself the experience for so long. It almost mystically transforms the act of spirited driving from a state of adolescent rebellion to a state of exhilarating civilization.

The 190's highly sophisticated multilink independent rear suspension system endows it with quick, keen, yet not overly anxious reflexes. And with a ride finely calibrated to give the driver an optimum feel for the road—but so substantial that potholes and other perils of the highway do not upset the 190 Class sedan's composure.

You are surrounded by a steel body structure welded into what one automotive journal terms "anvil-like solidity." The seat beneath you is built not to emulate a racing cockpit but to alleviate the fatigue of long-distance driving. You savor craftsmanship that led *Car and Driver* to ask, "How is it that Benzes fit together better than anything else in the world?" You relax in the knowledge that the Mercedes-Benz Supplemental Restraint System is primed to deploy a driver's-side air bag and front seat belt emergency tensioning retractors—within milliseconds of a major frontal impact.

You drive an automobile that affords ownership pleasures so satisfying that more Mercedes-Benz owners would buy again than would owners of any other make. Pleasures that may help explain why Mercedes-Benz automobiles as a line—and regardless of age—have retained a higher percentage of their original value than any other make.

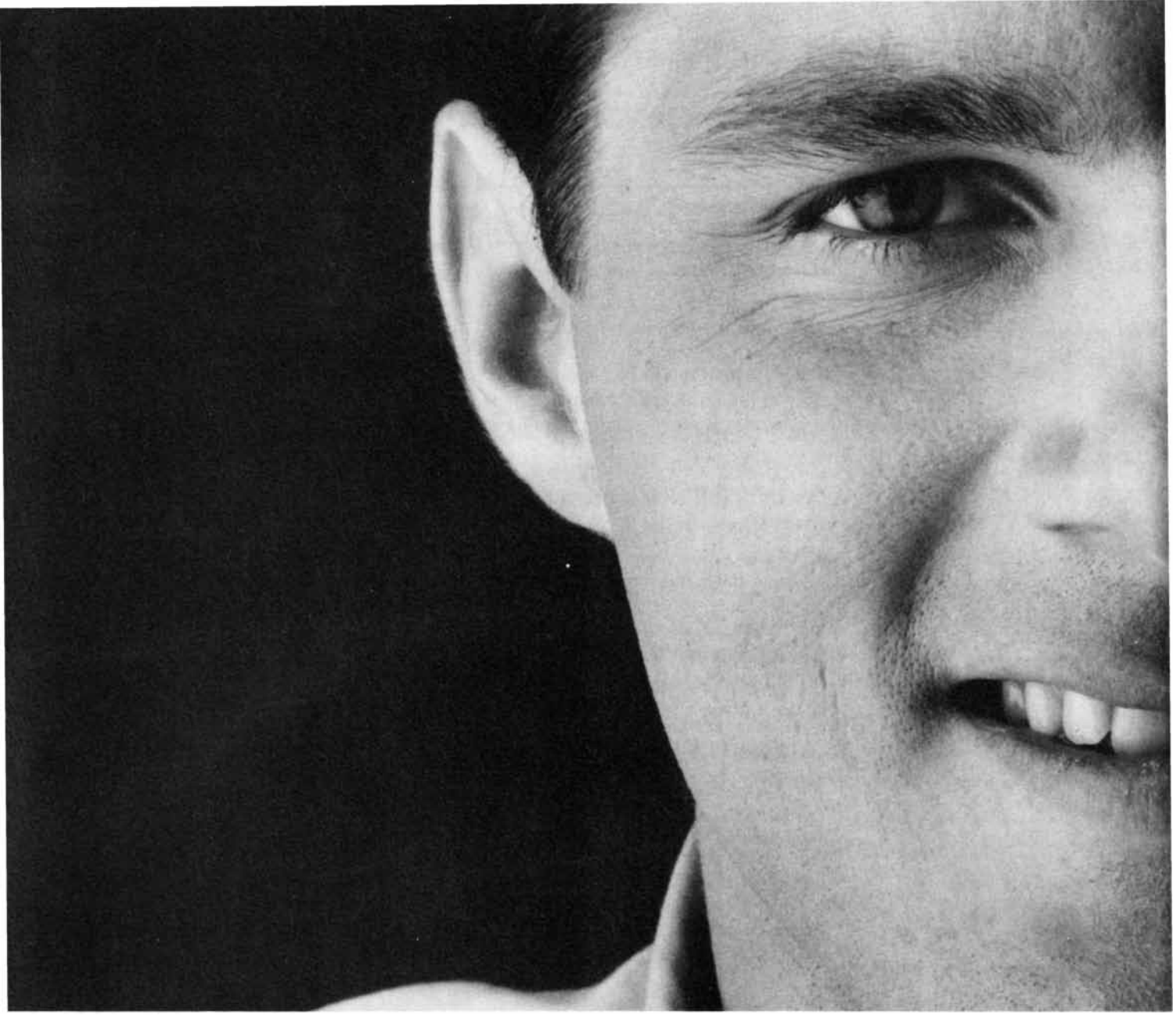
The Mercedes-Benz 190 Class, in brief, captures the most elusive of all engineering goals: that liberating sense of driving by mature and uncanny instinct. And rescues the sports sedan from its prolonged adolescence.

Engineered like no other car in the world

SCIENTIFIC AMERICAN

August 1988 Volume 259 Number 2

DEPARTMENTS



For the first time, users of UNIX® Syste

Announcing ETA System V.

Until today, if you used a system based on the real AT&T UNIX System V operating system, your lips were sealed when it came to working with a supercomputer. Making the move to true supercomputing meant having to learn a complicated proprietary operating system.

Those days are over. Because the first native operating system for a supercomputer based on AT&T's UNIX System V has arrived—ETA System V.

For the first time, users can work with the most powerful computers in the world using familiar commands.

ETA System V meets all the requirements of AT&T's System V Interface Definition (SVID) Release 3.0. It has passed all 5,500 tests in the System V Verification Suite. It is also the only supercomputer operating system to support features like ipc, semaphores and shared memory. Unlike non-standard operating systems, ETA System V has the advantages of byte addressability, virtual memory support, BSD sockets and r-commands.

ETA System V makes it possible to develop applications on industry-standard workstations compatible with AT&T's

m V can talk to a true supercomputer.

UNIX System V and then compile and run them on any ETA10 Supercomputer—from the affordable ETA10-P, the ETA10-Q and ETA10-E to the world's most powerful supercomputer, the ETA10-G.

Now, all your present applications based on AT&T's UNIX System V can be ported easily to ETA10 Supercomputers. Programs that once took months to port now take hours. Because programs based on AT&T's UNIX System V don't have to be rewritten.

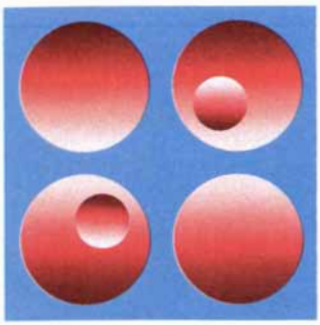
ETA is also developing a library of applications specifically for AT&T's UNIX System V users in higher education environments. These applications include SPSS, IMSL and DI3000.

If you know AT&T's UNIX System V, you now have a voice in supercomputing. Beginning today.

Talk to your Control Data representative, or: ETA Systems, Inc., 1450 Energy Park Drive, St. Paul, MN 55108. Phone: (612) 642-3460.

ETA10 SUPERCOMPUTERS.

CONTROL DATA



THE COVER demonstrates how the interpretation of shape from shading is powerfully affected by the human visual system's tendency to assume illumination from above (see "Perceiving Shape from Shading," by Vilayanur S. Ramachandran, page 76). The two small shaded disks superposed on the larger disks offer a particularly striking example: the right-hand disk suggests a sphere and the left-hand one a cavity. If the page is turned upside down, however, all the shapes reverse in depth.

THE ILLUSTRATIONS

Cover illustration by George V. Kelvin, Science Graphics

Page	Source
31	Julius Chang, State University of New York at Albany
32-35	Bob Conrad
36	James E. Vogelmann, University of New Hampshire (*left*); Ann Carey, U.S. Forest Service (*right*)
37-38	Bob Conrad
41-45	Ian Worpole
46	Jack D. Griffith, University of North Carolina School of Medicine at Chapel Hill
49	J. Bernard Minster and Thomas H. Jordan
50-56	Hank Iken
61	CERN, the European laboratory for particle physics (*top*); Gabor Kiss (*bottom*)
62-66	Gabor Kiss
69	Richard L. Edelson
70-74	George V. Kelvin
77	Vilayanur S. Ramachandran
78-79	George V. Kelvin
80	George V. Kelvin (*top*), Ron James (*bottom left*), George V. Kelvin (*bottom middle and right*)
81-83	George V. Kelvin
85	Gerald K. Skinner
86-88	Andrew Christie
89	Gerald K. Skinner
91	Tom Molesworth
92	Allan R. Mackintosh
93-96	George Retseck
100-103	Michael Goodman
105	Benno Hess, Max Planck Institute for Nutritional Physiology and the Stiftung Volkswagenwerk in West Germany
106	Laurie Grace
107	Courtesy of the estate of Fritz Goro (*left*), Benno Hess (*right*)
110	©1988, *Atlas of Blood Cells,* D. Zucker-Franklin, M. F. Greaves, G. E. Grossi and A. M. Marmont. Lea & Febiger

Scientific American (ISSN 0036-8733), published monthly by Scientific American, Inc., 415 Madison Avenue, New York, N.Y. 10017. Second-class postage paid at New York, N.Y., and at additional mailing offices. Authorized as second-class mail by the Post Office Department, Ottawa, Canada, and for payment of postage in cash. Subscription rates: one year $24, two years $45, three years $60 (outside U.S. and possessions add $11 per year for postage). Postmaster: Send address changes to Scientific American, Box 953, Farmingdale, N.Y. 11737-0000.

SCIENTIFIC AMERICAN

Established 1845

EDITOR: Jonathan Piel

BOARD OF EDITORS: Armand Schwab, Jr., Managing Editor; Timothy Appenzeller, Associate Editor; Timothy M. Beardsley; John M. Benditt; Laurie Burnham; Elizabeth Corcoran; Ari W. Epstein; Gregory R. Greenwell; John Horgan; June Kinoshita; Philip Morrison, Book Editor; Tony Rothman; Ricki L. Rusting; Karen Wright

ART: Samuel L. Howard, Art Director; Murray Greenfield, Associate Art Director; Edward Bell, Assistant Art Director; Johnny Johnson

COPY: Sally Porter Jenks, Copy Chief; M. Knight; Michele Moise; Dorothy R. Patterson

PRODUCTION: Richard Sasso, Vice-President Production and Distribution; Managers: Carol Eisler, Manufacturing and Distribution; Carol Hansen, Electronic Composition; Leo J. Petruzzi, Manufacturing and Makeup; Carol Albert; Nancy Mongelli; Jody Seward; William Sherman; Julio E. Xavier

CIRCULATION: Bob Bruno, Circulation Director; William H. Yokel, Circulation Manager; Lorraine Terlecki, Business Manager

ADVERTISING: Peter B. Kennedy, Advertising Director; Laura Salant, Sales Services Director; Diane Greenberg, Promotion Manager; Ethel D. Little, Advertising Coordinator

OFFICES: NEW YORK: Scientific American, 415 Madison Avenue, New York, NY 10017; Lisa Carden, Kate Dobson, Robert Gregory. CHICAGO: 333 North Michigan Avenue, Chicago, IL 60601; Litt Clark, Patrick Bachler. DETROIT: 3000 Town Center, Suite 1435, Southfield, MI 48075; William F. Moore. ATLANTA: Reese & Associates. CANADA: Fenn Company, Inc. DALLAS: Griffith Group. PRINCETON: William Lieberman, Inc. WEST COAST: Frank LoVerme & Associates

INTERNATIONAL: AUSTRALIA: International Media Reps, (02) 977-3377. FRANKFURT: Infopac, 6172-347-25. GENEVA: Infopac, 41-22-43-9435. HONG KONG/SOUTHEAST ASIA: C. Cheney & Associates, (852) 5-213671. KOREA: Biscom, Inc., (822) 739-7840. LONDON: Infopac, (441) 734-1343. PARIS: Infopac, 14-722-1265. SINGAPORE: Cheney Tan Associates, (65) 2549522. TOKYO: Nikkei International, Ltd., (13) 270-0251

BUSINESS MANAGER: John J. Moeling, Jr.

PRESIDENT OF MAGAZINE DIVISION AND PUBLISHER: Harry Myers

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

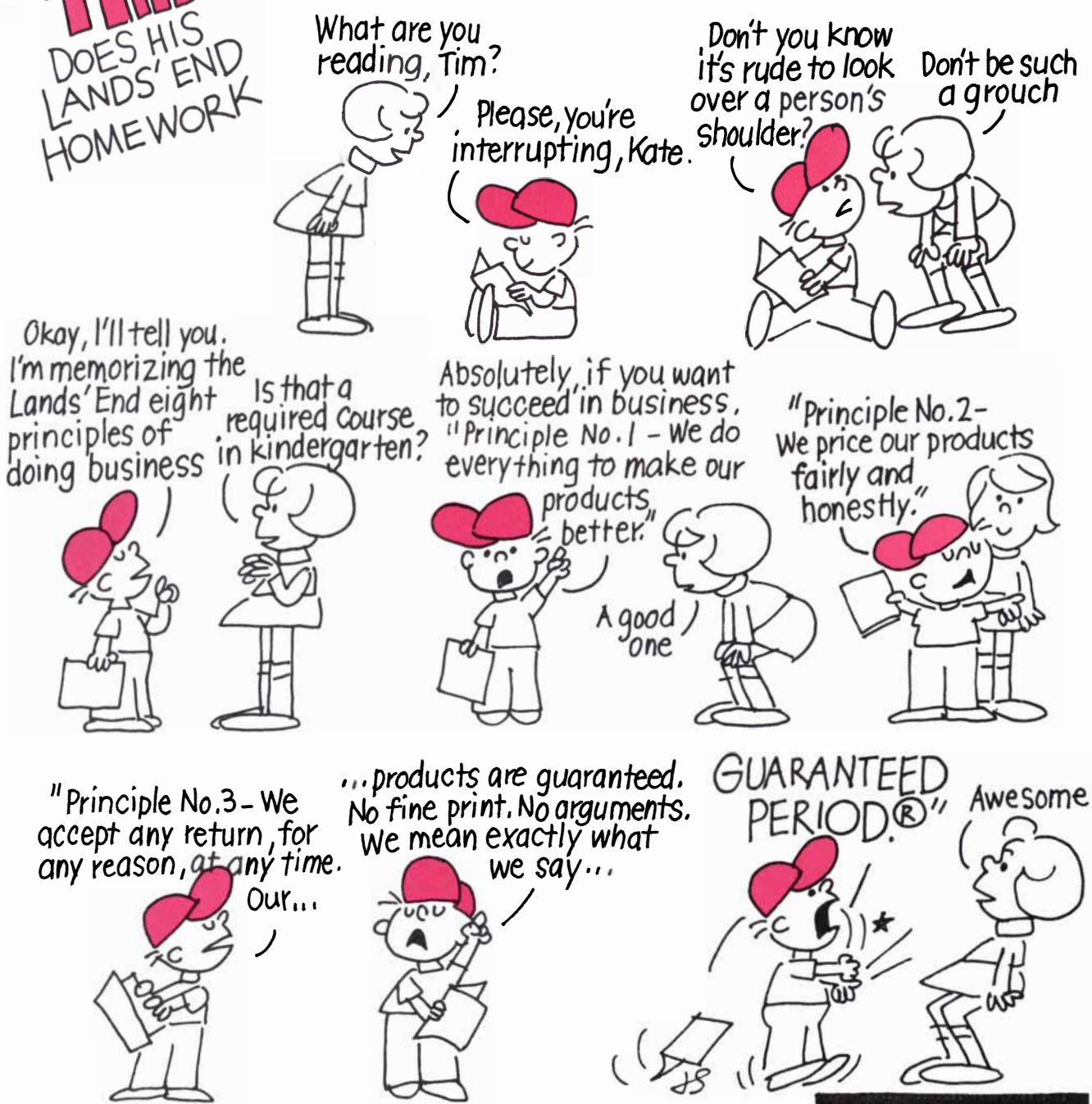
PRESIDENT AND CHIEF EXECUTIVE OFFICER: Claus-Gerhard Firchow

EXECUTIVE COMMITTEE: Claus-G. Firchow; Executive Vice-President and Chief Financial Officer, R. Vincent Barger; Senior Vice-President, Harry Myers; Vice-Presidents, Linda Chaput, Jonathan Piel, Carol Snow

CHAIRMAN OF THE BOARD: Georg-Dieter von Holtzbrinck

CHAIRMAN EMERITUS: Gerard Piel

Tim DOES HIS LANDS' END HOMEWORK



Principles 4, 5, 6, 7 and 8 are available, if you're interested. We happen to think our customers are the most important people in the world. That's what our 8 principles are all about. So, write or call for a catalog to check out our new Lands' End Children's Wear Collection and see what we mean by quality, value and service. Tim didn't make it up. Everything Lands' End sells is GUARANTEED. PERIOD.®

LANDS' END DIRECT MERCHANTS

Please send me a free catalog.
Lands' End Inc.
Dept. Q-F7, Dodgeville, WI 53595

Name _____
Address _____
City _____
State _____ Zip _____

Or call Toll-free
1-800-356-4444

LETTERS

To the Editors:

The photograph on page 20 of the June *Scientific American,* showing a cruise missile exploding over a target aircraft on San Clemente Island, is indeed spectacular. As a botanist attempting to preserve the 14 unique plant species and three unique animal species on the island, I get some bitter memories from that picture. The spectacle of the airburst was enhanced by dousing the craft with gasoline just before the launch. The mess made by the explosion still rests inside its $100,000 dirt bunker. The fields of white plywood panels used as ground references by the missile in navigating to its target are still scattered about the island, several of them in the middle of the habitat of the endangered San Clemente Island Sage Sparrow.

Many other senseless, destructive activities continue in the name of national defense on San Clemente Island, one of the more biologically unusual continental islands of the U.S.

R. MITCHEL BEAUCHAMP

Pacific Southwest Biological Services
National City, Calif.

To the Editors:

The caption for the photograph on page 17 of the March *Scientific American* states that soldiers from the U.S. and the U.S.S.R. both use "a heavy, rubberized suit" as part of their standard chemical-protection equipment.

According to the *Field Manual* of the U.S. Army, the U.S. has several styles of chemical-protection clothing. The standard-issue overgarment is permeable to air and moisture. Contrary to the assertion in the caption, this overgarment consists of an outer layer of nylon-cotton fabric and an inner layer of charcoal-impregnated polyurethane foam. Most U.S.S.R. overgarments are indeed made of rubberized fabric.

BRIAN T. COFFEY

Marquette Heights, Ill.

EDITOR'S NOTE

The DNA-amplification technique described in "Supertests" ("Science and the Citizen," February) was invented by Kary B. Mullis, formerly of the Cetus Corporation.

50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

AUGUST, 1938: "The mute and nerveless skeleton, which seems to the layman the one unvarying human constant, is actually never the same in two persons, nor for that matter does it remain unchanged for long in any given individual. It alters with our years and is a telltale index of our health, our way of life and frequently enough the manner of our death."

"The Zeppelin, dependent upon inflammable hydrogen for its lifting gas, has proved far too dangerous; now, with fireproof helium, it could again assume its place in aviation. As things now stand, Germany can build the airships but has no helium, while we can produce the helium but have no large airships in which it can be used. In view of the obvious need for airship service across the Atlantic Ocean between North America, Europe and South America, perhaps a compromise will be reached that will further the cause of lighter-than-air craft and will not create military complications."

"The highest pressure that up to now has been subject to laboratory control and accurate study is about 50,000 atmospheres; this is the pressure to be found in the crust of the earth at a depth of about 100 miles. About 92.5 percent of the material of the earth lies below a depth of 100 miles, so that our ignorance of earth conditions is still profound, but it is significant to have got this 7.5 percent under some sort of control."

"Quite accidentally the British have made the discovery that the metal structure of an airplane in flight collects and re-radiates or reflects the ultra-short-wave impulses employed in television broadcasting, so that receiving sets produce a double, or 'shadow,' image. The shadow image is formed by the waves that reach the television receiver directly and by those that rebound from the plane flying within range. It was also discovered, still by accident, that the width of the shadow image cast by the airplane reflections bears a definite relationship to the plane's distance."

"Two billion barrels of crude petroleum were saved to the world in 1937 by the chemical process of cracking heavy crude oil. Without this process the world would have required four billion barrels of oil to make the necessary gasoline, whereas the world production was only two billion."

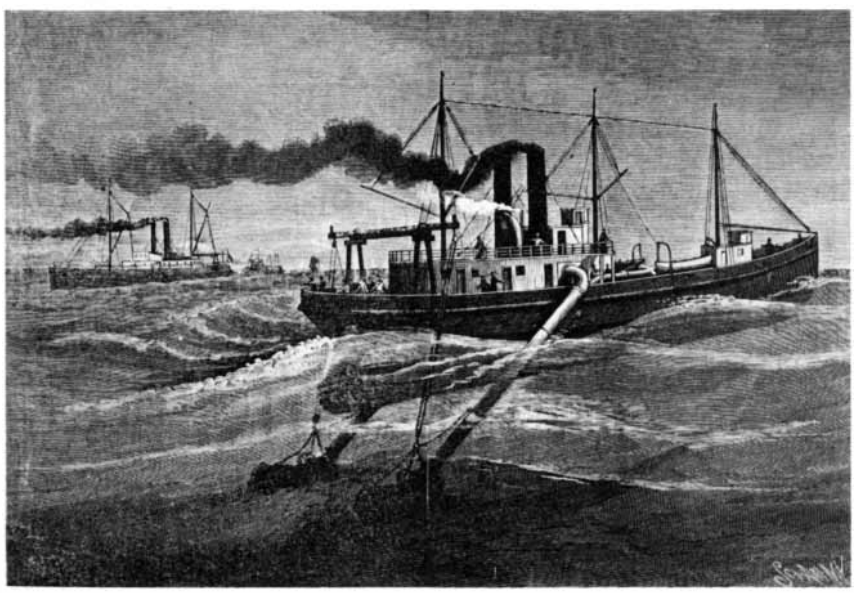
SCIENTIFIC AMERICAN

AUGUST, 1888: "The importation of fire crackers this year will amount to 300,000 boxes, an increase of 100,000 boxes over last year's importation. It is a little curious that the scientific knowledge and inventive genius of this country have proved inadequate for the successful manufacture of these explosives. All attempts to produce them in this country, so as to compete with the imported article, have failed. They are made in China and Japan, and the importation of the last week in June was 14,415 boxes, valued at $34,255. What a large sum to be thrown away on such trash!"

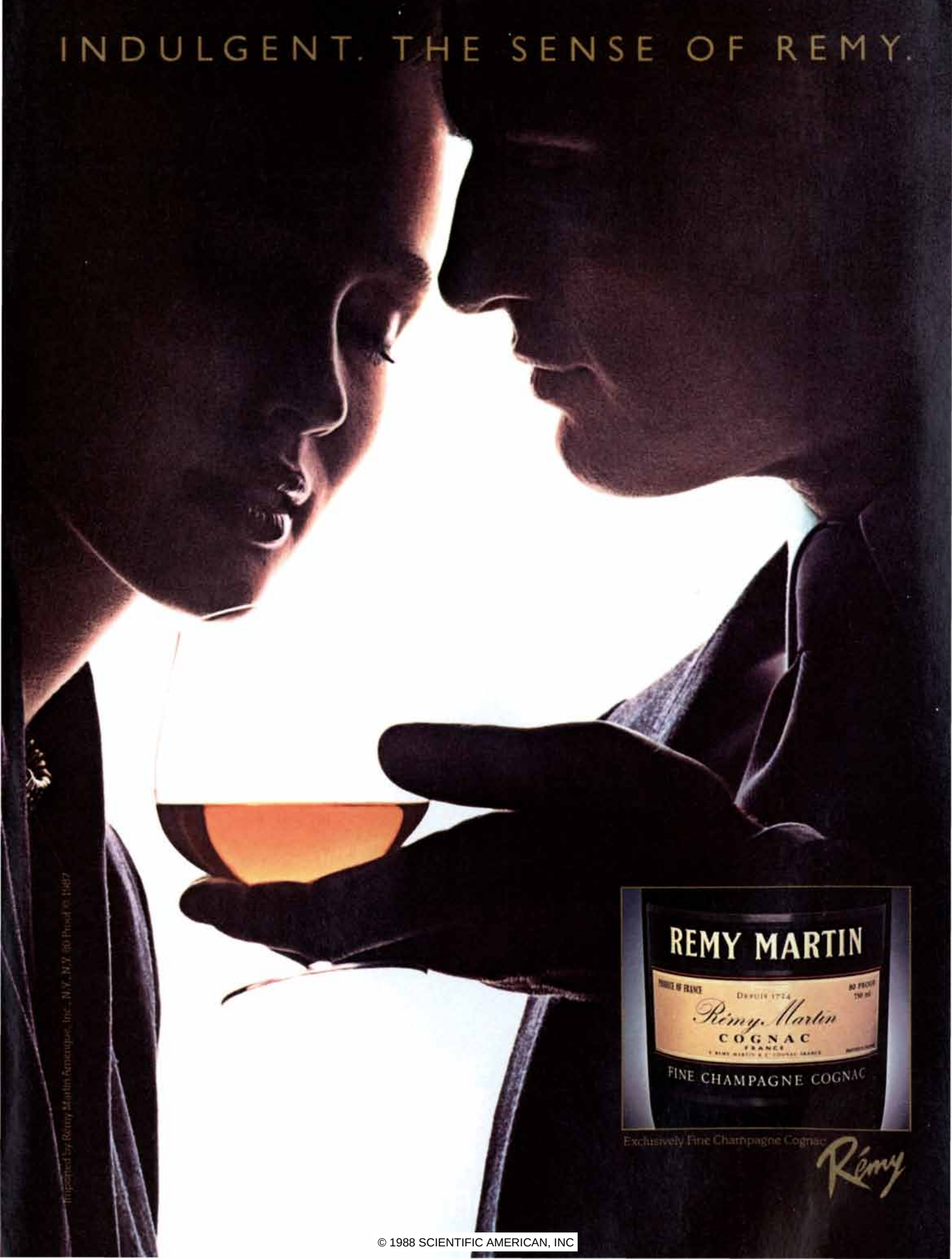
"The city of Reading, Pa., had a remarkable visitation of moths on the evening of August 1. Myriads of them infested the air, resembling at a distance a snow storm. They were first noticed flying around the electric lights about 8 o'clock, and gradually increased to such numbers as to obscure the brilliancy of the lights. Street-car passengers were covered with the insects, and handkerchiefs, hats, and fans were plied vigorously to keep them off. Local savants pronounced them cotton moths, and they evidently came from the South. They are said to precede a hot wave."

"Does anyone except a practical chemist ever stop to think of all the substances which we get from pit coal and the almost inconceivable variety of their uses? Everybody is familiar with those of them that are in daily use, such as gas, illuminating oils, coke, and paraffine, but of the greater part few persons know even the names. Among other things, there are obtained from coal the means of producing over four hundred colors, or shades of color, among the chief of which are saffron, violet blue, and indigo. There are also obtained a great variety of perfumes—cinnamon, bitter almonds, wintergreen, queen of the meadows, clove, anise, camphor, thymol (a new French odor), vanilline, and heliotropine."

"One of the most important operations ever conducted by the United States government, from a commercial point of view, is now in process of execution. We allude to the improvement of the channels leading up to New York City from the ocean. We illustrate the apparatus now in use by the Joseph Edwards Dredging Company, who are the sole contractors under this appropriation. The vessels are each fitted with two Edwards centrifugal pumps and two scoops connected by pipes with the pumps. Each vessel is divided into tanks for the reception of the dredged material."



Dredging vessel at work in New York Harbor



SCIENCE AND THE CITIZEN

B-2 or Not B-2

Stealth Bomber is the target in a debate on nuclear strategy

Late this fall, if all goes according to plan, the U.S. Air Force will publicly roll out what may become the most expensive weapon ever built: the B-2 Advanced Technology, or Stealth, Bomber. The aircraft, all details of which are highly classified, has been designed specifically as a strategic bomber that would enter Soviet airspace to make retaliatory strikes during a nuclear war. Every feature has been designed to make the aircraft hard for Soviet air defenses to detect. The result is a radical "flying wing" in which engines are embedded to make it hard for an infrared sensor to detect their heat. To reduce visibility to radar all external features have rounded contours, and much of the aircraft is built of carbon composites. The radar cross section, a measure of how easily radar can detect an object, is said to be one-thousandth that of a conventional airplane.

The B-2's political cross section is considerably more prominent. The Air Force acknowledges that the official estimate of $36.6 billion for a fleet of 132 B-2's has to be increased. Informed observers talk of a price tag of $400 million or more per plane. Estimates by the General Accounting Office put the total program cost for the B-2 at about $70 billion. Not all problems are budgetary. The prime contractor, the Northrop Aircraft Division, has had difficulty with manufacturing techniques it uses for some of the composites; program deadlines have slipped.

Spurred by spiraling cost estimates and other problems in the program, Robert Costello, Under Secretary of Defense for Acquisition, urged in June that the Defense Acquisition Board consider ending the program. One option that has been discussed is building a stealthy version of the B-1.

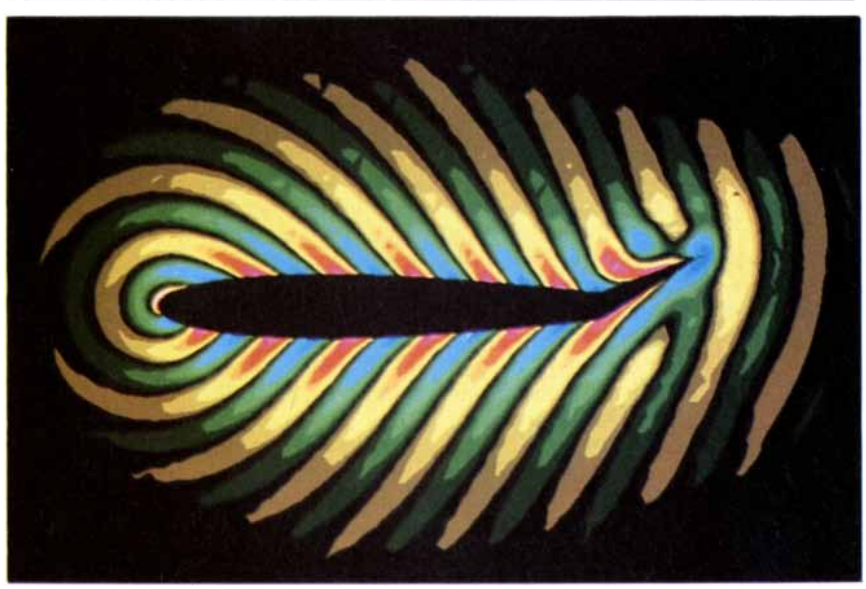
Although the B-2 continues to have support in Congress, that might be eroded if costs continue to rise. "It is seriously over cost, and we have to see whether we are getting the best bang for our bucks," says Representative Michael L. Synar of Oklahoma. On the other hand, Donald A. Hicks, former Under Secretary of Defense for Research and Engineering, argues that the program cost is comparable to that of the entire B-1 program, after allowing for inflation.

What does the U.S. get for its investment in the B-2? The primary strategic rationale offered by Hicks and such congressional supporters of the program as Senator Sam Nunn of Georgia and Representative Dave McCurdy of Oklahoma is that it will force the Soviet Union to spend more on air defenses (up to $400 billion by one estimate), thereby limiting Soviet resources for adventurism and offensive weapons. "That is the telling argument," says James R. Schlesinger, a former secretary of defense. Critics disagree. Herbert York, director of the Institute on Global Conflict and Cooperation at the University of California at San Diego, says the Soviet Union already spends $20 billion per year on air defenses. "They won't make any major changes one way or the other," he asserts.

Sidney D. Drell of Stanford University's Center for International Security and Arms Control and Thomas H. Johnson of the U.S. Military Academy have argued in *Foreign Affairs* that air-launched cruise missiles released from standoff bombers (ones that never enter Soviet airspace) could hit any target a penetrating bomber such as the B-2 could hit, and at far less cost and risk. Michael M. May, an associate director of the Lawrence Livermore National Laboratory, said of the B-2 at a recent congressional seminar that "it is difficult to find, so far at least, a serious deterrent mission for it within the nuclear strategic equation."

The Air Force contends that penetrating bombers might be able to search for mobile missiles such as Soviet SS-24's and SS-25's. Michael Brower of the Union of Concerned Scientists argues that target-recognition systems that might employ passive sensors (which do not emit radiation)

Some defense analysts in the Government think spending on the B-2 should be reduced—or ended



***RADAR BEAM** bathes a metal airfoil in a computer simulation made at the Rockwell International Science Center by Vijaya Shankar. Colors show the variance in intensity of the scattered electromagnetic radiation. Such studies help designers to reduce the radar reflectance of aircraft, making them "stealthier."*

will not be available for "many years, if at all." Likewise, May argues that in order to search actively for mobile missiles a stealth bomber would have to turn on its radar—and so become visible to defenses. Drell argues further that threats to mobile missiles are destabilizing in times of crisis.

In spite of the strategic doubts, supporters of the B-2 say it would be useful in missions other than nuclear attacks on the Soviet Union. Hicks and Schlesinger conjecture that it would have an invaluable role in dealing with a Soviet incursion into Europe or the Middle East; the B-2 compares well with the cost of maintaining an aircraft-carrier fleet, according to Hicks. The critics have a ready answer: to use the B-2 for anything other than its intended purpose, namely a nuclear bombing mission, would expose the secret technology to unreasonable risks of being captured. The debates are likely to continue. —*Tim Beardsley*

Report from Stockholm

Some progress is being made, but HIV still outwits therapy

John Benditt, a member of the Board of Editors, attended the Fourth International AIDS Conference in Sweden. His report follows.

There was not any news of breakthroughs for the 6,600 scientists and 700 journalists who packed the gleaming Älvsjö conference center on the southern edge of this lovely water-threaded city in mid-June. Instead, in a sober and businesslike mood that belied the brilliant sunlight and lingering twilight of *midsommar,* the conferees absorbed reports that added up to a deepening but still far from complete understanding of the complex life cycle of HIV, the AIDS virus. Such knowledge may ultimately provide new points of attack for antiviral therapies. It was the consensus of the conference that safe and effective therapy will be available before a vaccine—perhaps long before. In the absence of a vaccine, the conference was warned, rapid behavioral changes are needed to prevent the virus from spreading by way of young members of urban minority groups.

Perhaps the best overview of the life cycle of HIV was provided by William A. Haseltine of the Dana-Farber Cancer Institute. At one of the large daily plenary sessions of the conference Haseltine announced the discovery in his laboratory of a new HIV gene. The newly identified DNA segment, called *vpu* (for viral protein *u*), brings to nine the number of genes known to be involved in regulating the replication of the AIDS virus. Interactions among the array of regulatory genes, Haseltine said, could yield three different levels of virus replication: zero, a low level of chronic replication and a replicatory burst.

The possibility of a prolonged zero state of virus replication early in infection was "one of the chief concerns to have emerged from this meeting," according to Haseltine. Until now it has generally been thought that the initial phase of HIV infection was marked by a burst of virus replication and the simultaneous production of antibodies that can be detected by standard antibody tests. Results presented at the conference by Steven Wolinsky of the Northwestern University School of Medicine, however, show that at least in some individuals viral DNA is integrated into host-cell chromosomes long before the virus replicates.

Wolinsky described a sample of 18 men drawn from the large Multi-Center AIDS Cohort Study (MACS) initiated by the National Institute of Allergy and Infectious Diseases. Employing a new technique called the polymerase chain reaction (PCR) that amplifies specific viral-DNA sequences, Wolinsky was able to find viral DNA in cells from 16 of the 18 men prior to the appearance of antibodies as detected by standard methods. The average interval between infection and the appearance of antibodies was some 18 months, but in one case the gap was 42 months. It is not known how widespread the phenomenon of long-delayed antibody appearance may be, and studies are now being designed to answer that question, according to John P. Phair of Northwestern, one of the principal investigators of the MACS study.

Although these data are disturbing, they are not without reassuring aspects, Wolinsky noted in an interview. First, he said, the PCR technique offers "a means of providing early diagnosis and thereby counteracting the spread of HIV infection." In addition, his observations raise the question of how some infected people are able to control the virus completely for as long as 42 months. A study based on blood samples drawn from the MACS cohort is under way to answer that question. Finally, the capacity to detect small amounts of viral DNA will contribute to more accurate assessments of the efficacy of antiviral therapies.

Much attention was focused at the conference on antiviral strategies. The agents discussed ranged from the sole drug now approved for clinical use in the U.S.—AZT—to approaches that are still speculative. A report from Robert Yarchoan of the National Cancer Institute confirmed earlier indications that the effectiveness and safety of AZT therapy can be enhanced by combining the drug with a related one called dideoxycytidine (ddC). Both AZT and ddC are analogues of the nucleotides that are subunits of the DNA chain; they work by interrupting the synthesis of viral DNA. A problem with both agents has been their toxicity: AZT kills rapidly dividing blood cells, and ddC has been linked to painful neurological effects. Yarchoan reported that a broad trial has now shown that an experimental regimen in which AZT and ddC were alternated is effective and that it reduces the side effects of both compounds.

Yarchoan also expressed enthusiasm for a drug called dextran sulfate, which blocks infection of cells by HIV in the test tube, as Hiroaki Mitsuya of the NCI reported. Dextran sulfate, which has previously served as an anticlotting agent, has been tested in preliminary safety trials in AIDS patients by Donald Abrams of the San Francisco General Hospital. The drug was tolerated with little toxicity. "Dextran sulfate is exciting for several reasons," Yarchoan said, "not only because it is well tolerated and has a high therapeutic index in the test tube but also because it may be able to block the formation of syncytia." These are giant clusters of abnormal cells that some investigators believe are implicated in the pathogenic effects of HIV.

A sobering final note was provided on the last day of the conference by King K. Holmes of the University of Washington School of Medicine, an expert in sexually transmitted diseases. Holmes argued that if the epidemic of HIV infection is not checked, it could run a course parallel to that of other sexually transmitted diseases such as gonorrhea, which have attained high levels in some urban minority populations, particularly among young women. What is more, Holmes said, it is known that any disease causing genital ulceration increases the probability of HIV transmission. Therefore, he concluded, any anti-AIDS program omitting treatment for other sexually transmitted diseases is a "blueprint for disaster." He called for "targeting women at the earliest possible age, particularly women of low socioeconomic status," in programs that include drug counseling, treatment

for sexually transmitted diseases and job training. It is ironic, Holmes noted, that in recent years funds for the treatment of other sexually transmitted diseases have been diverted to the fight against AIDS.

K.A.L. 007

Did the U.S. "misrepresent" a key piece of evidence?

Nearly five years after a Soviet jet destroyed Korean Air Lines flight 007, questions still linger. One question not yet raised in the many public examinations of the tragedy concerns the integrity of a crucial piece of evidence: a tape recording of Soviet pilots intercepting the airliner.

White House officials released the recording on September 6, 1983, six days after the shoot-down. The officials never explained exactly how they acquired the recording (they implied that it came from a Japanese signals-intelligence post) or why it lacked transmissions from the ground-based controllers with whom the Soviet pilots were conversing. But the Administration's interpretation of the recording was unequivocal. According to Jeane J. Kirkpatrick, then the permanent representative of the U.S. to the United Nations, it proved that the Soviet pilots made no attempt to identify or warn the Korean jet before one of them shot it down. (The U.S.S.R. had claimed otherwise.) Kirkpatrick also announced before playing the tape at the UN: "Nothing was cut from this tape. The recording was made on a voice-actuated recorder and, therefore, it covers only those periods of time when conversation was heard."

A spectral analysis of the recording suggests it may have been neither voice-actuated nor unedited, according to Lawrence L. Porter, an investigator of aviation accidents who specializes in audio analysis. Porter, who spent 22 years with the Federal Aviation Administration before becoming a private consultant, analyzed the recording for the Fund for Constitutional Government. This nonprofit watchdog group, which was founded by liberal philanthropist Stewart R. Mott in 1974 and is based in Washington, had obtained a tape of the recording from the American mission at the UN.

Porter's suspicion centers on frequent pauses in the recording that, ostensibly, represent points at which the recorder shut off in response to a lull in the Soviet transmissions. The pauses last for only a fraction of a second and so are inaudible to an untrained ear. On a spectrogram, a plot of the frequencies of the various sound waves that make up a sound track, the pauses show up clearly as blank spaces above a thin band of low-frequency noise.

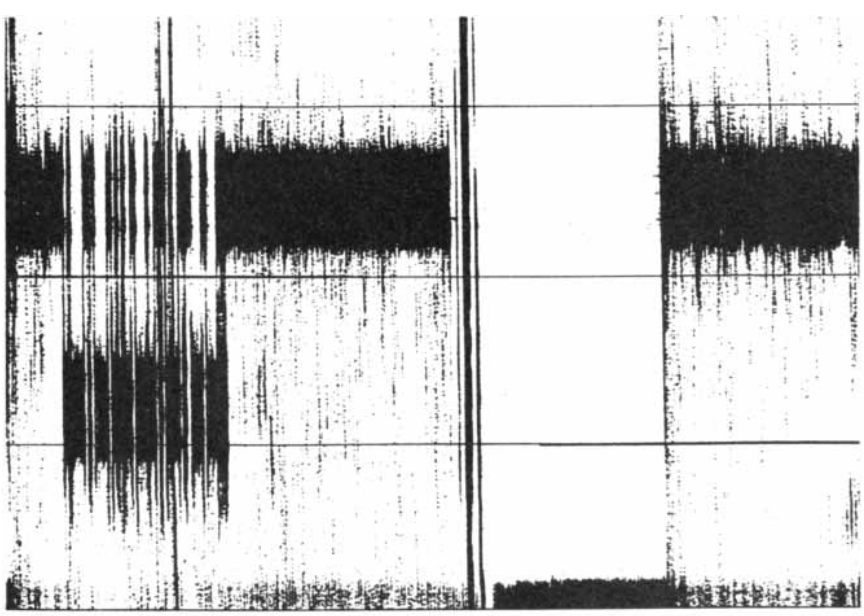
According to Porter, these pauses seem to differ from ones normally produced by a voice-actuated recorder in at least five respects: (1) Some of the pauses are not immediately followed by the sound of a voice, as one would expect of pauses in a voice-actuated recording. (2) The blank spaces produced by the pauses on the spectrogram have borders that are unusually sharp; they show no signs of the small transient signal normally produced by a voice-actuated recorder turning off and on. (3) The pauses vary in length; ordinarily the pauses on a voice-actuated recording are identical in length. (4) A high-frequency carrier tone visible as a dark band running across the top of the spectrogram intrudes into some of the pauses; the carrier tone, since it was part of the original Soviet transmission, should stop when the recorder stops. (5) One of the pauses is preceded by an audible click, visible on the spectrogram as a double bar; such a mark resembles one produced by editing.

Porter is also puzzled by the presence on the tape of brief oscillating tones that recur, for the most part, at one-second intervals. "There is no reason for them to be there," he says. He speculates that whoever made the recording added the tones to mark the passage of time. But, he adds, if that is the case, the intervals between the tones should not vary in length; actually at several points on the tape the intervals between the tones do vary, suggesting material may have been removed or added.

Porter interprets his findings cautiously. He says he cannot be certain the recording was deliberately edited to portray the Soviets' actions in the worst possible light. "The words I like to use are inconsistency and misrepresentation," he says. "This tape is not a continuous recording, as it was purported to be."

Charles M. Lichenstein, who was deputy U.S. representative to the UN Security Council in 1983, helped Kirkpatrick to prepare her dramatic presentation of the recording at the UN. He says the U.S. Information Agency supplied Kirkpatrick and him with tapes of the recording and told them it was voice-actuated and unedited. "We had no reason to think any misrepresentation was made to us," says

A tape of Soviet pilots chasing K.A.L. 007 was said to be unedited, but an analysis suggests otherwise



SPECTROGRAM of a recording of Soviet pilots pursuing K.A.L. 007 shows a click **(vertical bar)** ***followed by a pause*** **(empty region).** ***The dark band near the top of the spectrogram represents a carrier tone. A brief oscillating tone*** **(left)** ***periodically interrupts the recording. No speech is detectable in this segment.***

Lichenstein, now a senior fellow at the Heritage Foundation. He suggests, however, that USIA technicians may have edited the recording to "speed it up between voice bursts" and to "get rid of background noise."

John Keppel, manager of the Fund for Constitutional Government's K.A.L. 007 investigation, says Porter's work adds to other findings that suggest U.S. officials may have altered or withheld evidence related to the tragedy. Keppel, who served in the State Department for 22 years, says his own participation in the Eisenhower Administration's effort to conceal the true nature of a U-2 spy plane shot down over the Soviet Union in 1960 taught him "not to believe everything our Government says." He thinks Congress should reopen an inquiry into the 007 affair. Senator Edward M. Kennedy apparently agrees. According to his staff, he is asking the Permanent Subcommittee on Investigations of the Senate Governmental Affairs Committee to examine the role of the U.S. in the tragedy. —*John Horgan*

What Price Cost Controls?

Hospital belt tightening may be linked to poorer care

Are government attempts to cut the costs of medical care resulting in poorer treatment for patients? There is abundant anecdotal evidence—stories of particular patients who have received inadequate treatment because of cost-cutting policies—but there is not much hard data on the subject. Now a study published in the *New England Journal of Medicine* shows that in some cases there has indeed been a correlation between strict government-mandated cost controls and poor outcomes for patients: for certain groups of patients, hospitals in states that imposed tight regulations had higher mortality rates than hospitals in states with looser regulations.

The authors of the study, Stephen M. Shortell and Edward F. X. Hughes of Northwestern University, considered two kinds of cost controls, known respectively as rate-review programs and certificate-of-need programs. Under rate-review programs, hospitals are not allowed to increase their yearly revenues by more than a certain percentage each year. The percentage differs from state to state, as do the level of enforcement and the harshness of the penalties that are imposed for exceeding the limit.

Under certificate-of-need programs, hospitals are prevented from spending more than a certain amount of capital on such projects as added beds, new technology or expanded programs; they can exceed the limit if they first obtain a certificate of need justifying the expense. In states that have certificate-of-need programs, the dollar limit on capital expenditure and the difficulty of acquiring a certificate of need vary from state to state.

Shortell and Hughes examined data on Medicare patients receiving care for any of 16 selected conditions in 981 hospitals in 45 states. (They studied Medicare patients because an extensive data base is maintained for patients within the Medicare system.) Having controlled for such variables as the size of the hospitals, their mix of cases and the median incomes of the counties where they were situated, the workers found that the ratio of actual to predicted death rates of hospitals in states with stringent rate-review programs was from 6 to 10 percent higher than that of hospitals in states with less stringent programs. The ratio of actual to predicted death rates of hospitals in states that have strict certificate-of-need programs was from 5 to 6 percent higher than that of hospitals in states with less strict programs. "This is significant," notes Hughes, "because it corresponds with what clinicians have been saying for years."

Do these results mean that cost controls are killing patients? Not necessarily. The study shows a correlation, not a strict cause-and-effect relation; it reveals nothing about why the mortality rates differed between states with strict cost-control measures and states with loose ones. As Shortell says, "There are a lot of things we cannot conclude from our study. It is a first step, not a last step."

He believes the investigation's chief function has been to point out areas where further research is needed. In this study the hospital was viewed as a "black box": the investigators considered only the overall regulatory limits and the final outcomes of hospital care. Now it is time, says Shortell, "to look inside the black box. What are hospitals doing to control costs? We need to develop better monitoring systems to look at patient outcomes and the quality of care."

Shortell adds that such an examination will raise even deeper questions: How can the quality of care be measured? What constitutes an appropriate level of care? Assuming that some cost-control measures can be effected without harming patients, what is the threshold level at which a further reduction in cost inevitably leads to worse care? —*Ari W. Epstein*

PHYSICAL SCIENCES

Complexity Counted?

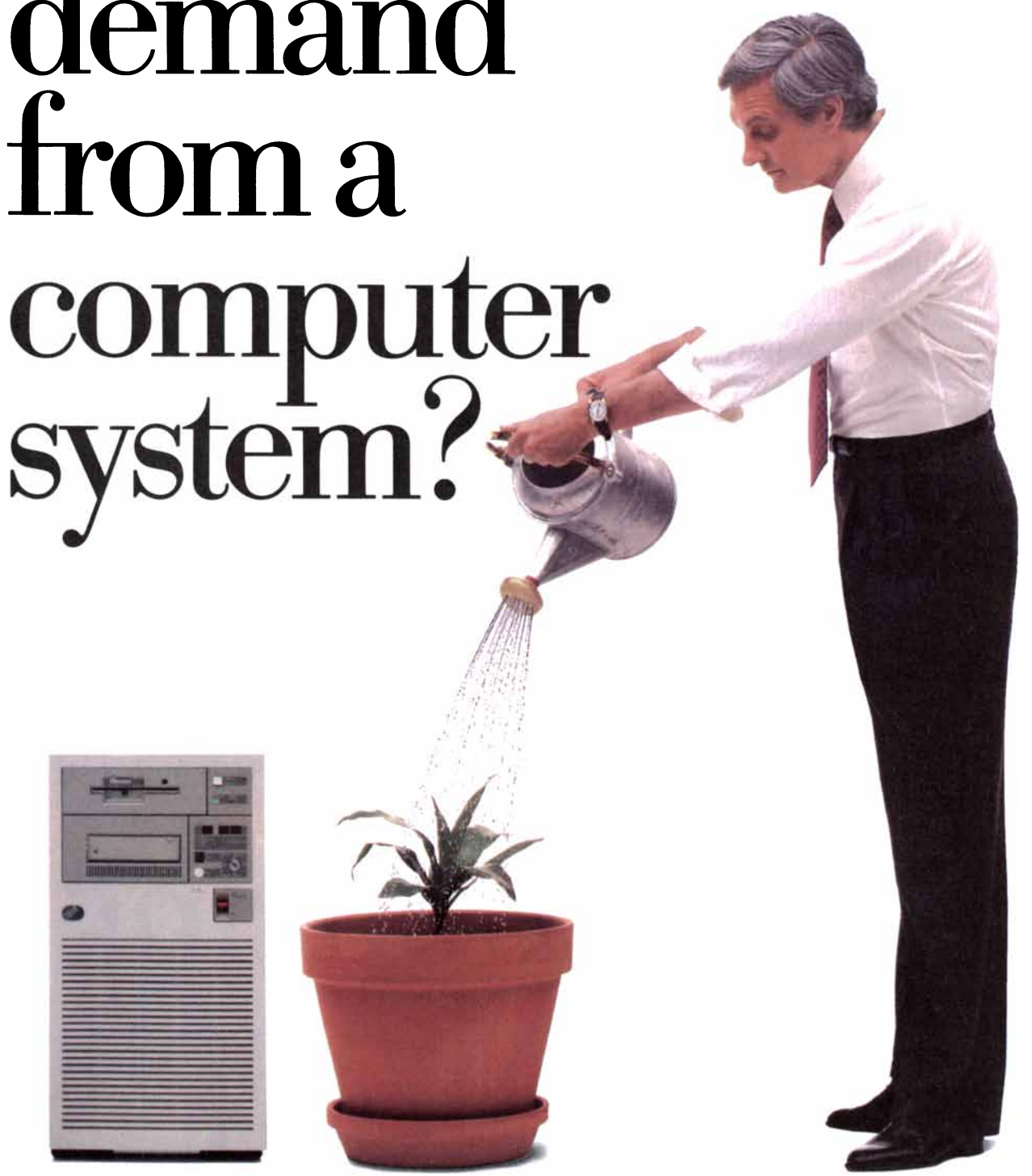
Physicists ponder a new way to measure an elusive concept

Most people would agree that a rose is more complex than a gas. But how much more complex? And is a rose more or less complex than a fruit fly? Various ways of measuring the complexity of numbers have been proposed in recent decades, but there is as yet no generally accepted measure of the complexity of a physical object.

A possible approach has been outlined by Seth Lloyd and Heinz Pagels of Rockefeller University, who have devised a measure called thermodynamic depth. The measure links complexity with thermodynamics. Lloyd explains that thermodynamic depth was formulated so that it would be zero for totally ordered states, such as the regular array of atoms in a diamond, and also for totally random states, such as the molecules in a gas. It would be high for intermediate states. Another requirement was that attaining complexity should not be easy: two bacteria should be considered less than twice as complex as a single bacterium, since bacteria can make copies of themselves easily.

These requirements and a few other stipulations led Lloyd and Pagels to a concept based on the process by which an object is created. The thermodynamic depth of making a car from scratch, for example, is equal to the thermodynamic depth of making all the parts from scratch plus the thermodynamic depth of putting them together. Technically the present thermodynamic depth of a system is the difference between two quantities. The first is its entropy (a measure of the observer's lack of exact knowledge about the system). The value of the second quantity depends on the amount of information needed to specify all the paths by which the system might have reached its present state from its measured state at some earlier time. There are many possible paths, because the system might originally have been in any one of many different but quite indistinguishable states.

What should every company demand from a computer system?



Growth.

When your business is small, you can buy an IBM® Application System/400,™ and it will be just the right size.

Later on, you'll still be smiling.

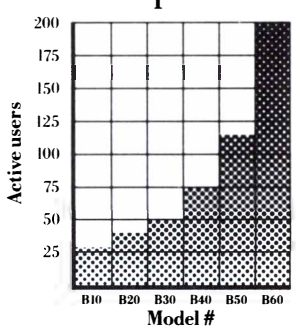
That's because as your company grows, your Application System/400 can grow right along with you. And the investments you made at first—in software, training and peripherals—will still be working for you.

That's what the IBM Application System/400 is all about. It comes from IBM's leadership with over a quarter million mid-size computer systems in place, and it does what growing companies have told us they want.

It lets you grow into what you need, without outgrowing what you've paid for.

Today: Solutions for your business, from the leader in business solutions.

Never before has a mid-size computer system been introduced with so much proven software ready to go. Thousands of programs that run on IBM's System/36 and /38 can run on the IBM Application System/400.



However big your IBM Application System/400 may get, the same software will always work.

Even better, any program you start with, you can stay with. No matter how big your IBM Application System/400 gets, your software will always work.

Tomorrow: Protection for your investment as your business grows.

Most models of the IBM Application System/400 come rack-mounted like a stereo system. Components slide in and out, so you can upgrade right on the spot. You're not locked into any particular setup. Your system's only as big as you need, and making it bigger is easy.

What's more, the Application System/400 comes with a state-of-the-art education system, plus customer support that's unmatched in the industry.

For a free brochure, or to arrange for a detailed discussion about the Application System/400, call 1-800-IBM-2468, ext. 80.

Introducing the IBM Applica

tion System/400.

IBM The Bigger Picture

More intuitively, the thermodynamic depth of a process is a measure of how difficult it is to assemble something: the difference between the amount of information needed to describe the system now and the amount needed to describe all the states it might have been in at the start of the process. The thermodynamic depth is consequently proportional to the amount of information the process has discarded.

The measure has some satisfying features. A bacterium has a large thermodynamic depth because evolution has discarded great amounts of genetic information over the aeons in arriving at the present-day bacterium. The additional thermodynamic depth incurred when it makes a copy of itself is relatively small. The measure also encompasses earlier definitions of complexity. Lloyd says the response of physicists who have read a preliminary account of his and Pagels' work is "guardedly enthusiastic," although some people also say that the arguments have yet to be made rigorous and that the measure has some undesirable properties.

What use is a measure of complexity? Lloyd thinks that if thermodynamic depth is found to be generally applicable, it could be a tool for studying complex systems in general, including evolution and such biological processes as the self-assembly of proteins, where the universal tendency of matter to become more disordered is locally reversed. He says thermodynamic depth might also elucidate notions that have until now only been conjectures, such as the proposition that complex systems are necessarily thermodynamically unstable. Lloyd is applying the measure to explore the idea that the evolution of complex systems might be inevitable. —*T.M.B.*

Going for Gold

What does "black stuff" say about the origin of oil and gas?

A 22,000-foot experimental drill shaft in Sweden recently unearthed 60 kilograms of "extremely smelly" black sludge from granitic rock. The "black stuff," as it has been called, has the consistency of modeling clay and incorporates what seem to be molecules of biological origin. How it got there has become the focus of a sharp dispute between supporters and critics of a maverick theory about the origin of oil and gas.

The theory is that of Thomas Gold, professor of astronomy at Cornell University. The Swedish State Power Authority, the Gas Research Institute in Chicago and private investors are drilling the hole on the edge of the Siljan ring, an ancient meteorite crater, to test Gold's ideas. Most geologists think oil and gas are the chemically transformed remains of organisms buried under ancient seas. Gold believes they result instead from bacterial action on hydrocarbons that were trapped deep within the earth's mantle at the time of its formation and are slowly seeping toward the surface. If Gold is right—and so far few geologists are convinced he is—the earth's reserves of oil and gas would be vastly greater than is now thought.

Gold thinks the black stuff is evidence supporting his theory, since conventional wisdom does not predict that oil or gas would be found in granite at such depth. The material consists of fine-grained magnetite (an oxide of iron) and various hydrocarbons as well as "biomarkers," or carbon compounds characteristic of biological activity. The biomarkers constitute a chemical fingerprint almost identical with that of oil found in some nearby surface seeps. The pungent stink suggests recent bacterial activity, according to Gold, who sees a

possible analogy with bacteria associated with deep ocean vents. Bacteria are known to be capable of making magnetite. Gold notes too that there are anomalously high levels of iridium in the black stuff. Iridium is also present in significant amounts in some meteorites that may have contributed hydrocarbons to the young earth.

In addition to the black stuff, increasing quantities of various hydrocarbon gases were found as the drilling progressed, as well as hydrogen and helium. Gold argues that contaminants, such as lubricants put down the shaft, cannot be the source of the hydrocarbon gases or the black stuff, as critics have suggested. The main reason he gives is that levels of helium, which is not present in any of the drilling additives but is known to emanate from the deep earth, correlate well with measured levels of the other gases and so suggest a common origin for the gases.

Alan Jeffrey of the Global Geochemistry Corporation in Los Angeles, on the contrary, thinks biological marker molecules from oils near the surface somehow leached down the drill hole and accumulated there. He points out that the black stuff was found inside the hollow drill pipe, not outside it as one might expect. Robert Hefner III, an Oklahoma gas prospector who is receptive to Gold's ideas, counters that strange pressure effects are known that could explain how the material was sucked into the drill pipe.

Paul A. Westcott of the Gas Research Institute thinks the helium measurements could be a result of occasional pressure drops in the shaft. Such drops tend to make any gases present come out of solution, and so they could explain the apparent correlation with other gases. He and Jeffrey both think the hydrocarbons could have come from oils added as lubricants during drilling.

The project recently faced financial difficulties, but the Swedish State Power Authority is sufficiently convinced to have decided to resume drilling. The current plan is to go down to about 24,500 feet. —*T.M.B.*

A Match Made in Heaven

U.S.-Soviet collaboration in space science is improving

On a table within his office at the National Aeronautics and Space Administration, Samuel W. Keller, who manages collaboration with the Soviet Union, displays a volume of radar maps of Venus obtained in 1983 by the Soviet *Venera 15* and *Venera 16* orbiters. In return for the Venus data, Keller recently delivered to Moscow the first installment of a set of computer-enhanced maps of Mars, made by reanalyzing data gathered by the U.S. Viking missions in the 1970's and early 1980's. The maps will help Soviet space scientists to select touchdown sites for a Mars lander they hope to launch in 1994. The exchange is a tangible symbol of the renewed collaboration in space science brought about by the warming of U.S.-Soviet relations.

Many of the joint plans reflect the current fascination with Mars. The Soviet missions to the Martian moon Phobos, scheduled to be launched in July, will be the first test of the new cooperative spirit. The two probes will orbit Mars and, after their rendezvous with Phobos, will release two types of landers and remotely analyze surface rocks by detecting ions blasted off them with a laser and a plasma gun. If the first probe is successful, the second one may be diverted to the other moon, Deimos. As the probes orbit Mars the U.S. will follow their progress with the tracking antennas of its Deep Space Network and will provide the Soviet Union with accurate data

on their positions, according to Keller.

Collaboration is extending to the spacecraft themselves. Under agreements reached at the recent summit meeting in Moscow, spacecraft from one country will soon be carrying instruments from the other country. Until a few months ago it appeared unlikely that the Reagan Administration would ever clear U.S. instruments for flight on Soviet spacecraft, for fear of revealing technology with military applications. The U.S. interagency group that reviews export proposals, however, recently agreed that two U.S. space technologies—albeit relatively unsophisticated ones—could be flown on Soviet spacecraft. Keller says the move is "probably a major step forward."

One device is a 1960's-vintage Total Ozone Monitoring System (TOMS), an instrument similar to the one that, flying on a U.S. weather satellite, *Nimbus-7,* provided decisive evidence of a worldwide decrease in stratospheric ozone. The hiatus in the American space program means the U.S. will not launch a replacement for the instrument before its predicted failure date; the Soviets have the necessary launch capacity. The second technology approved for launch on a Soviet rocket is an imaging component for a U.S.-Danish X-ray telescope.

The instrument exchange will work both ways. Keller informed the U.S.S.R. Academy of Sciences in Moscow that the U.S. is willing in principle to carry a Soviet radio receiver on the *Mars Observer* mission, an orbiter that is scheduled for launch in 1992. The receiver would serve as an additional way of relaying signals from the 1994 Soviet Mars lander back to the earth.

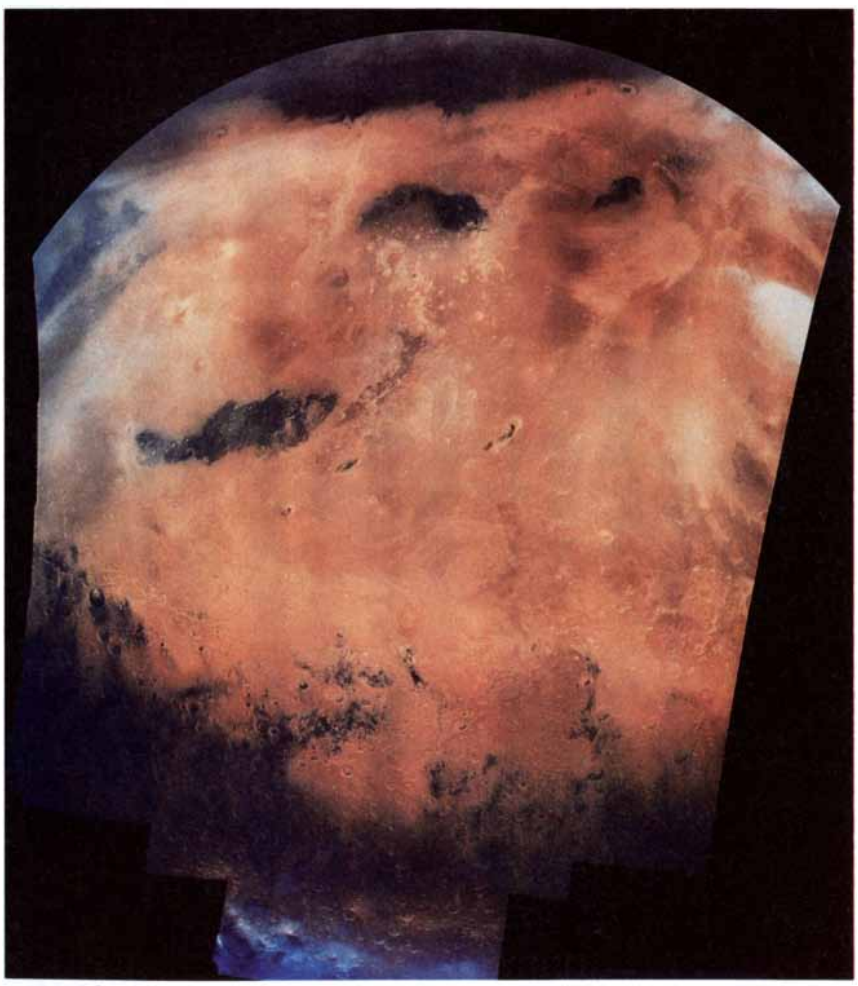
Many people have been urging more ambitious collaborations. The biggest questions about Mars—whether it has ever supported, or even still does support, any kind of life—will probably not be firmly answered until rocks have been returned to the earth for analysis. Soviet space scientists have suggested that an unmanned mission to gather samples and return them to the earth could be second in the lineup of missions after their 1994 lander (which the Soviet government has not yet formally approved).

Organizations such as the Planetary Society and an informal association of Mars enthusiasts known as the Mars Underground are pushing hard for a U.S. commitment to a joint manned mission to Mars. Congress is also enthusiastic: the House of Representatives recently went so far as to pass a bill directing NASA to start planning by 1992 for such a mission.

The Administration is definitely not ready for such a step. "None of us believes it is timely to make a commitment to a manned Mars mission," according to Keller. He suggests that several years of lesser collaboration and study should precede a decision on whether to cooperate on such a massive and uncertain venture. The Administration remains wary even of joint unmanned Mars missions; at the Moscow summit the U.S. declined Soviet General Secretary Mikhail S. Gorbachev's overtures about such a mission, agreeing instead only to separate national studies of possible areas of collaboration in space science.

In the meantime U.S. investigators, whose own programs are stagnating, must content themselves with vicarious involvement in Soviet efforts. Harold Masursky of the U.S. Geological Survey in Flagstaff, Ariz., who is reanalyzing the Viking Mars data, would like his computer-enhanced mosaic maps to lead the Soviet Mars lander to sites with complex local geology and rocks of various ages. He says image enhancement has enabled him to discern valleys cut by long-vanished rivers in lava flows from at least four successive epochs. He therefore believes there was water on Mars as recently as a few hundred million years ago, which might improve the chances that Mars once hosted life. (Traditional wisdom holds there has been no water on Mars for billions of years.) The truth may be known by the end of the century, but, on the basis of missions

Computer-enhanced maps of Mars made in the U.S. could help the U.S.S.R. to choose a landing site



VIEW OF MARS was composed from 104 digital images of the planet made by* Viking 1 *in 1980. The mosaic was made by Jody Swann of the U.S. Geological Survey with funding from the National Aeronautics and Space Administration.

"Accords have traditionally been some of the nicest cars to sit in.

They are also satisfying on the move. The engine seems to hum sweetly. Cruising is hushed. The steering, deliciously accurate. And the shifter carves to perfection.

The Honda Accord is world-class comfortable."

–Car and Driver

It's enough to make us blush.

The Accord LX HONDA

discussed so far, it is likely to be a Soviet spacecraft that brings back the crucial evidence. —*T.M.B.*

BIOLOGICAL SCIENCES

Truth or Consequences?

How should institutions handle charges of misconduct?

Scientists are not usually audited: it is assumed that investigators will avoid making fraudulent claims and even doing sloppy work because of the likelihood that they will be exposed when other workers cannot duplicate their findings. Yet recent, widely publicized accusations of scientific misconduct have convinced many people that research institutions do not always respond adequately to such allegations.

The frequency of scientific misconduct, which can include anything from careless reporting practices to fraud, or deliberate falsification, is far from clear. On the one hand, Daniel E. Koshland, Jr., the editor of *Science,* has written that "99.9999 percent" of published scientific reports are accurate. On the other hand, data gathered by the Food and Drug Administration, which does audit clinical trials of experimental drugs, present a less rosy picture. Eleven percent of 1,758 effectively random audits done by the agency over 10 years turned up evidence of significant misconduct. Since 1977 the FDA has disqualified or negotiated restrictions on 68 investigators. Fourteen investigators have been convicted on such charges as knowingly misleading the Government, and some of them were fined or given jail terms.

The consequences of fraud can be serious. A former researcher at the University of Pittsburgh, Stephen E. Breuning, was indicted in April for providing false information to the Government in support of a grant application. A review organized by the National Institute of Mental Health had concluded earlier that Breuning fabricated data about the effects of drugs on mentally retarded children. According to the NIMH review, the data influenced public-health policy in several states.

Yet most universities lack experience in organizing investigations of misconduct charges, even though they generally have some formal policy for doing so. A recent spate of bruising inquiries has starkly illustrated the cost of foot dragging and inadequate protection for both the accuser and the accused.

The National Institutes of Health last year upheld several charges of scientific misconduct against a researcher at the Cornell University Medical College after an investigation that had lasted for more than five years and cost the accuser $13,000 of his own funds; Cornell's inquiry had previously dismissed the charges. Earlier this year a committee in the House of Representatives chaired by Representative John D. Dingell of Michigan held a fiery public hearing at which a former postdoctoral fellow at the Massachusetts Institute of Technology testified that she was forced out of her scientific career because she questioned the interpretation of experiments given in a paper published by some of her colleagues.

The former fellow, Margot O'Toole, had raised questions about a paper on immunology that was published in 1983 by Thereza Imanishi-Kari (now at Tufts University), Nobel laureate David Baltimore (the director of the Whitehead Institute for Biomedical Research at M.I.T.) and others. Informal reviews impaneled by M.I.T. and Tufts found no compelling reason to think O'Toole's alternative interpretation of the experiments was correct, but two unofficial investigators of misconduct working at the NIH, Ned Feder and Walter Stewart, subsequently pursued O'Toole's allegations. Congress has now begun its own inquiry. Baltimore has agreed, in a letter circulated to his colleagues, that the paper contained one error, but he strongly denied any wrongdoing. He wrote that the paper "appropriately reflected the state of the science at the time it was written."

Expressing a common sentiment about the policing of research, Baltimore also warned that the congressional investigation might serve to spur new regulations that "could cripple American science." It was on similar grounds that the Office of Management and Budget recently blocked new regulations, proposed by the Public Health Service, that would specify procedures for universities faced with allegations of misconduct. The OMB argued that such regulations should address only charges of fraud, because attempts to define misconduct might stifle scientific creativity.

Yet Robert M. Andersen of the National Science Foundation, which last year introduced regulations prescribing how institutions that receive NSF funds should deal with allegations of misconduct, points out that fraud itself is usually impossible to prove. Andersen believes Federal agencies should have a broader mandate: to intervene if host institutions do not investigate accusations of misconduct promptly and fairly.

Meanwhile many universities are reexamining their procedures for investigating charges of misconduct, and the Institute of Medicine is studying how misconduct might be discouraged. Members of the study panel are said to be particularly concerned that junior researchers in medical schools are inadequately supervised. The IOM may also recommend that the practice of "honorary authorship," whereby the names of senior investigators are added to a study's list of authors to increase its luster, be discouraged; the practice is thought to undermine the responsibility of authors to ensure the accuracy of studies published under their names. —*T.M.B.*

A Better Crystal Ball

A new predictor of diabetes suggests a means of prevention

A specific autoantibody (an antibody against the self) has been found to be an excellent predictor of insulin-dependent (type I) diabetes. The finding could lead to a new test for identifying people who will develop the disease, which arises when the insulin-producing beta cells of the pancreas are destroyed, probably by an autoimmune process. The finding may also point the way to preventing overt disease in the identified individuals.

The investigators—Mark A. Atkinson, Noel K. MacLaren and William J. Riley of the University of Florida College of Medicine and David W. Scharp of the Washington University School of Medicine—knew about the autoantibody from earlier work by a Danish group. It is directed against an antigen with a molecular weight of 64,000 daltons that is normally found on the surface of beta cells. To evaluate how well the autoantibody (which has come to be known as the 64K autoantibody) predicts future diabetes, the workers screened blood samples that had been drawn from 12 randomly chosen diabetics long before the subjects developed the disease. Remarkably, the 64K autoantibody was found to have been present in all 12 subjects from three months to seven years before the onset of diabetes, suggesting that the autoantibody (which did not appear in any healthy subjects) is highly predictive of future diabetes.

SCIENCE/SCOPE®

The feasibility and benefits of adapting the Free Electron Laser (FEL) to perform precision radar measurements in space will be examined under a new Hughes Aircraft Company program. Scientists will study the potential of the FEL as a compact space-radar transmitter that would be part of a Strategic Defense Initiative (SDI) system for discriminating objects in space. The program will take advantage of the FEL's inherent tunability, high power and efficiency, and its ability to operate in frequency bands of 100 GHz and higher. The program's ultimate goal is a space-based, multiband, adaptive FEL capable of operating efficiently at randomly chosen, stable frequencies.

A new orbiting sensor system could potentially serve as a space-borne missile warning system. Under development for the Strategic Defense Initiative Organization (SDIO), the Boost Surveillance and Tracking System (BSTS) satellite is designed to maintain continuous surveillance of Earth. Hughes will develop the system's infrared sensor and signal processor, which will be able to provide reliable detection of hostile missiles, even in severe countermeasure environments. The design will be developed to incorporate an efficient modular growth path to meet all of the SDIO's future requirements for the BSTS satellite.

A new technique for interconnecting high lead-count integrated circuit (IC) chips to a substrate or package is being developed as a software and hardware package in a single piece of equipment. Called single-point Tape Automated Bonding (TAB), the Hughes system uses an etched tape, typically formatted in 35-mm or 70-mm sizes, on which a specific IC pattern is etched out. A polyimide film is used to separate and support the leads used for the interconnects. Single-point TAB combines the speed and precision of tape-automated bonding with wire bonding's ability to handle a wide variety of chip shapes, including the latest very large-scale integrated chips.

A new Dome Display System will incorporate background and other target projectors with twice the resolution and twice the targets of those previously simulated by other trainers. The system, provided by Hughes, will be used in Lockheed's YF-22A Advanced Tactical Fighter (ATF) prototype development program. The display will include a 28-foot dome and other equipment similar to that used in Hughes' F/A-18 Weapons Tactics Trainers. The Dome Display System is one of several ATF programs currently being developed by Hughes.

Hughes' Santa Barbara Research Center has openings for qualified applicants experienced in Infrared Systems and Program Engineering activities. You will be responsible for developing design verification plans, preparing radiometric performance predictions, and applying advanced technology to Infrared Sensor Systems. If your experience includes infrared sensor design, detector technology, and low noise analog circuits, contact the Santa Barbara Research Center, Professional Employment, Dept. S2, 75 Coromar Drive, Goleta, CA 93117. EOE. U.S. citizenship required for most positions.

For more information write to: P. O. Box 45068, Los Angeles, CA 90045-0068

HUGHES
AIRCRAFT COMPANY

Subsidiary of GM Hughes Electronics

Atkinson and his colleagues, who reported the data at the annual scientific sessions of the American Diabetes Association, found that the 64K autoantibody appeared either before or at the same time as two other autoantibodies that often appear in prediabetics. He says this indicates that the 64K autoantibody may well be the earliest reliable marker of impending diabetes.

The early appearance of the autoantibody suggests that the 64K antigen to which it responds may help to initiate the autoimmune destruction of beta cells. One piece of evidence in support of this idea is the fact that the 64K autoantibody is seen only in people who have diabetes or in whom diabetes is developing, never in other people. In addition, Atkinson and his collaborators at Florida and in Denmark have detected the 64K antigen only on beta cells, not on any of 15 other cell types tested so far. An autoimmune process that destroys only beta cells would presumably be triggered by an antigen that appears on those cells and no others.

If the 64K antigen is indeed involved in the autoimmune process, the finding would raise the possibility of developing a treatment to block its effects. No one has yet identified the initial trigger for diabetes, but it is possible that a virus carrying an antigen similar to the 64K antigen invades the beta cells and provokes both a normal immune response to the virus and also an abnormal, autoimmune response to the 64K antigen. Once that has happened, beta-cell destruction could proceed in at least two ways. The 64K autoantibodies could bind to the antigen and stimulate other parts of the immune system to destroy the bound beta cells. Alternatively, immune-system cells known as *T* lymphocytes could recognize the antigen and destroy the cells directly. In either case, Atkinson says, a therapy he and his colleagues hope to test in animals within a year or two has the potential to interfere with both processes, thereby delaying or preventing insulin-dependent diabetes.

In the proposed therapy, which would have to continue for life, the workers would inject into the bloodstream a toxin bound to a purified form of the 64K antigen. The antigen-toxin complex would quickly reach the lymph nodes, where it would be taken up by immune cells that normally produce the 64K autoantibody. Through a more circuitous route the complex would also be bound by the *T* lymphocytes that recognize the 64K antigen on beta cells. Thus the specific cells involved in beta-cell destruction would themselves be poisoned and inactivated, leaving nondestructive cells unharmed. Cancer researchers are already testing similar "immunospecific" approaches to killing tumors.

The roadblock to testing this treatment is the fact that the sequence of amino acids constituting the 64K antigen is not known. That information is crucial to the development of both the proposed therapy and an inexpensive screening test for the 64K autoantibody. Atkinson plans to devote most of the coming year to determining the sequence. *—Ricki Rusting*

So Long, Lefty

Slightly sinister statistics for left-handed people

Left-handed people may have more to worry about than just the global scarcity of left-handed scissors, golf clubs, fishing reels and other necessities. A study described in a recent letter to *Nature* suggests they have a slightly shorter life expectancy than right-handers.

The lack of reliable data on the handedness of dead people had forestalled such studies until Diane F. Halpern of the California State University at San Bernardino and Stanley Coren of the University of British Columbia came up with an ingenious solution: *The Baseball Encyclopedia.* That volume records which hand the late greats (and the not-so-greats) of baseball favored when they threw and batted, as well as their dates of birth and death.

The two workers analyzed a sample of 1,472 righties and 236 lefties. (Oddballs such as switch-hitters or those who threw from one side and batted from the other were excluded.) Measured year by year, the mortality rates for the two groups are practically identical up until the age of 33. After that the left-handers' mortality rate exceeds that of the right-handers by an average of 2 percent for most ages. Another statistic seems somewhat less alarming: the mean life span for the left-handers was 63.97 years, only about eight months less than the mean life span for the right-handers.

Halpern and Coren (both of whom are right-handed) had suspected they might arrive at such findings. Some years earlier Coren had participated in a study that found 13 percent of a group of 20-year-olds were left-handed but only 5 percent of a group in

their fifties. Other studies have indicated that left-handers are more likely than right-handers to suffer from allergies and other immunological disorders. There is also evidence that left-handers are accident-prone, particularly when using cars with stick shifts on the right and other equipment designed for the majority.

Halpern acknowledges that the recent study's sample population is "rather unusual," since it consists entirely of men who survived at least into early adulthood and were (at that point) exceptionally athletic and healthy. She and Coren would like to gather data from a broader population, including children and women. They are considering working with a large random sample of the recently deceased, whose handedness they would establish by mailing a questionnaire to close relatives. —*J.H.*

TECHNOLOGY

Son of Rubber

Rubber doped with iodine becomes an electrical conductor

Nothing is the way it should be anymore. Dinosaurs are warm-blooded, television evangelists are embroiled in scandals and rubber conducts electricity. That, at least, is the surprising claim of Mrinal K. Thakur of the AT&T Bell Laboratories, who reports in *Macromolecules* that natural rubber treated with iodine becomes a conductor.

Some polymers are known to conduct electricity when they are doped with particular chemicals (see "Plastics That Conduct Electricity," by Richard B. Kaner and Alan G. MacDiarmid; SCIENTIFIC AMERICAN, February). That natural rubber should behave in the same way is unexpected. The known conducting polymers all have one feature in common: the carbon atoms that form the backbone of the molecule are connected by alternating double and single chemical bonds; this "conjugated bond" structure makes the polymer susceptible to doping. Furthermore, conducting polymers tend to be black, rigid and insoluble.

Undoped natural rubber, or polyisoprene, is transparent, is not rigid and is soluble in hexane. Most important, it does not share the conjugated-bond structure that is characteristic of the other conducting polymers; its carbon atoms are all connected by double bonds.

Conventional wisdom held that the conducting electrons in doped polymers travel along the chain of conjugated bonds. Last year, however, it was shown that in polyacetylene heavily doped with iodine, electrons hop from one chain to another; the finding has resulted in a debate on the dominant mode of conduction.

The polyacetylene results gave Thakur the idea of doping rubber with iodine. He hoped that the iodine would pull an electron out of the double bond to make a "charge-transfer complex" capable of conducting across the polymer chains. Thakur's experiment was simplicity itself: he placed a thin film of rubber in a vacuum chamber and sublimated iodine onto the rubber.

Thakur found that the conductivity increased by about 12 orders of magnitude, from 10^{-13} siemen per centimeter (characteristic of a good insulator) to 10^{-1} siemen per centimeter. (This is still three orders of magnitude less than the conductivity of iodine-doped polyacetylene and seven orders of magnitude less than the conductivity of copper.) Moreover, it seems that the conductivity is indeed achieved by electrons hopping across the polymer chains.

Thakur remarks that it is too early to predict applications, "but the material is cheap and easily obtainable, so somebody will come up with something." —*Tony Rothman*

Robots Rampant

California artists spawn technological monsters

Survival Research Laboratories. Sounds like another high-technology startup, a purveyor of futuristic weapons perhaps. Actually this San Francisco–based organization does deal in complicated machines with a destructive bent. But the goal is not profit. It is art, performance art.

"We're trying to develop a theater that revolves around machines," says Mark Pauline, an expert welder and machinist who founded SRL about a decade ago. Working primarily with hardware scavenged from the region's many defunct factories and machine shops, Pauline's troupe has created such marvels as the One-Ton Walking Machine, which resembles a skeletal elephant; the Big Arm, a cross between a backhoe and a dinosaur, and the Inspector, which looks like a terribly uncomfortable hospital bed equipped with long, clawed arms.

During an SRL performance, usually staged in a parking lot or other damage-resistant arena, the gasoline- and diesel-powered robots crawl, stagger and hurtle into one another to the accompaniment of a cacophonic sound track. Pauline and other SRL members usually lurk offstage, controlling the machines with radio transmitters. Some of the robots can also operate autonomously. A guinea pig encaged by a set of contact switches once piloted a flame-throwing walking machine. The Big Arm has recently been given a more conventional onboard brain: a programmable microprocessor.

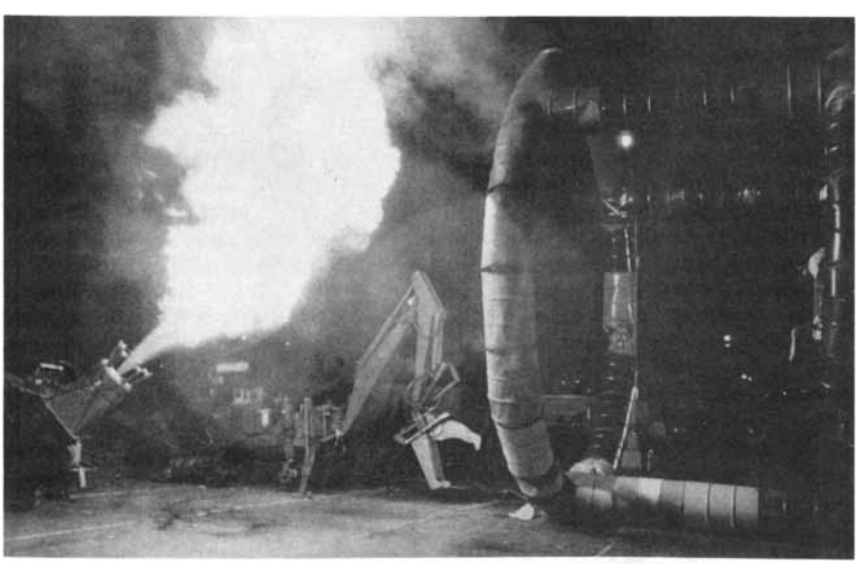
SRL has won a following not only among avant-garde aesthetes but also among engineers, some of whom lend their expertise to the troupe. Phillip H. Paul, a mechanical-engineering researcher at Stanford University who has followed SRL's progress for some seven years, helped to design one of its noisiest "special effects" devices. Called the Shock-Wave Cannon, it focuses the explosion of an oxygen-acetylene mixture into a shock front that can shatter glass 100 feet away. "What impresses me most about SRL is their ability to tackle some pretty tough problems in a reasonable amount of time and at no cost," Paul says. He notes that machines such as the One-Ton Walking Machine, although relatively "crude and heavy," do essentially the same things that robots built for millions of dollars by Government and industrial researchers do.

Rick Rees of Bell Northern Research, who helped to design the Big Arm's computer-based control system, suggests that SRL fills a persistent void in modern culture. "Artists and engineers usually don't speak the same language," he says. "SRL is blazing new ground by trying to build a collaboration between art and technology."

Pauline professes dislike for most art that incorporates technology. Too often it "serves the status quo of the art world," he says. "The art world wants something very packaged and ordered." SRL shows are not very packaged and ordered. This was evident during a show one rainy night last spring in a parking lot outside Shea Stadium, in Flushing, N.Y. At one point the Walking Machine bumped into the two-story-high Big Wheel, a low-tech but dangerous-looking contraption made of oil drums welded together, and sent it rumbling into a light pole. SRL technicians rushed out and heaved the Big Wheel away from the swaying, sputtering light. Then the Big Wheel lurched toward the Shock-Wave Cannon. A fiery blast from the Sprinkler from Hell, an industrial sprinkler turned flamethrower, blistered the paint on the Big Wheel but failed to stop it. Finally the Big Arm seized the Big Wheel and stopped it just short of the Shock-Wave Cannon. The audience, soaked and shivering, shrieked its approval.

Asked later about the incident, Pauline said, "We planned it." —*J.H.*

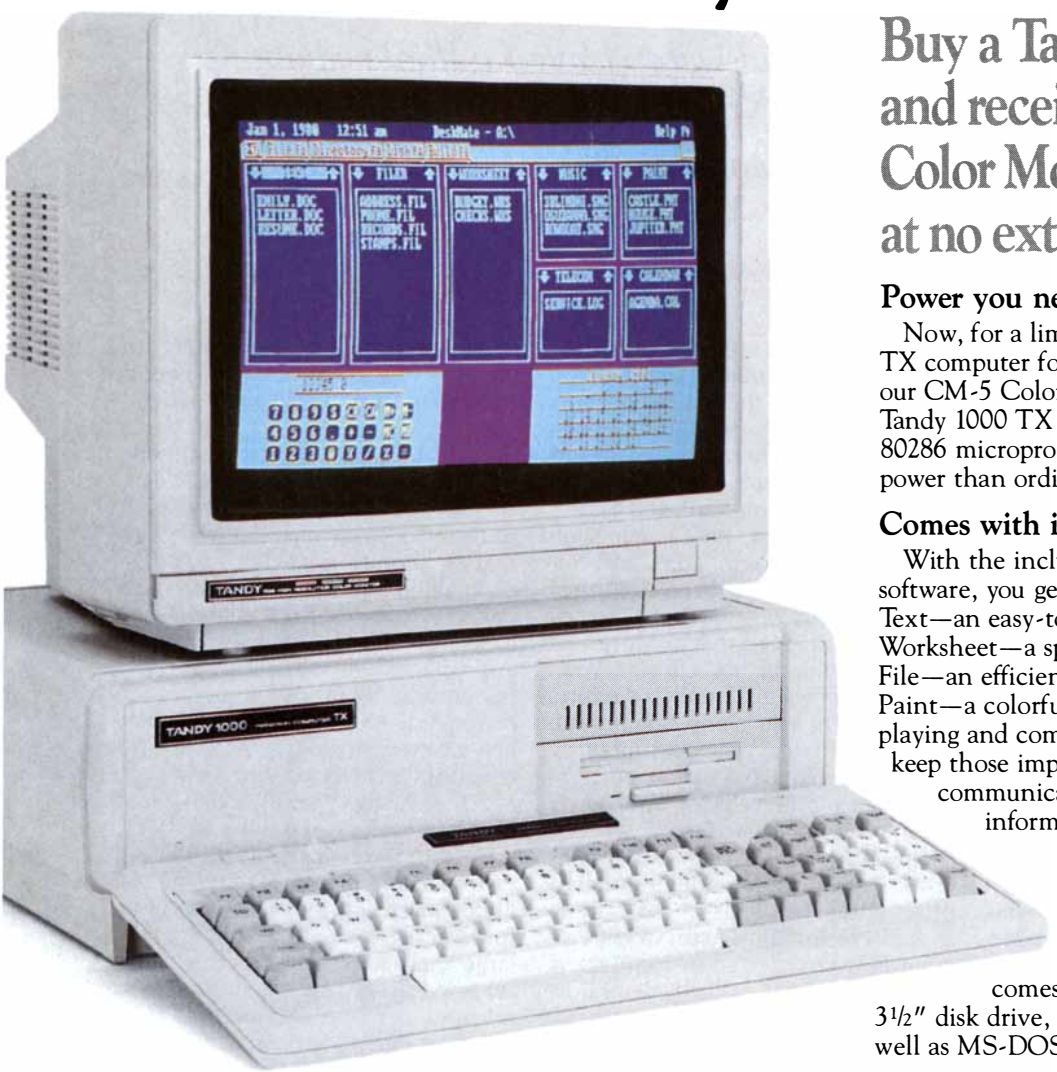
Artist-mechanics transform cast-off machinery into the stars of an inhuman, violent theater



***SPRINKLER FROM HELL** scorches the Big Wheel as the Big Arm looks on during an SRL performance in Flushing, N.Y. Photograph by Bobby Adams.*

Tandy Computers: Because there is no better value.™

The Tandy® 1000 TX



Buy a Tandy 1000 TX and receive a $299.95 Color Monitor... at no extra charge.

Power you need at a great price

Now, for a limited time, buy a Tandy 1000 TX computer for only $1199 and we'll include our CM-5 Color Monitor. The PC-compatible Tandy 1000 TX features a high-speed Intel® 80286 microprocessor for far greater processing power than ordinary PCs.

Comes with its own software

With the included Personal DeskMate™ 2 software, you get seven popular applications: Text—an easy-to-use word-processing program; Worksheet—a spreadsheet-analysis application; File—an efficient electronic-filing system; Paint—a colorful graphics program; Music—for playing and composing songs; Calendar—to keep those important dates; and Telecom—to communicate with other computers and information services.

Start computing immediately

This system is ready to run from day one because the TX comes with 640K RAM, a 720K 3½″ disk drive, all the necessary adapters, as well as MS-DOS® 3.2 and GW-BASIC.

Choose from a variety of computers

Tandy offers a complete line of PC-compatible computers for every need. Visit a nearby Radio Shack today and take advantage of this special offer featuring the remarkable Tandy 1000 TX with Personal DeskMate 2 and the CM-5 Color Monitor.

Radio Shack®
COMPUTER CENTERS
A DIVISION OF TANDY CORPORATION

Offer includes Tandy 1000 TX (25-1600) and CM-5 Color Monitor (25-1043). Monitor appearance may vary. Personal DeskMate 2 communications require modem. Intel/Reg. TM Intel Corp. IBM/Reg. TM IBM Corp. MS-DOS/Reg. TM Microsoft Corp. Sale begins 6/21/88, ends 8/23/88.

SCIENTIFIC AMERICAN August 1988 Volume 259 Number 2

The Challenge of Acid Rain

Acid rain's effects in soil and water leave no doubt about the need to control its causes. Now advances in technology have yielded environmentally and economically attractive solutions

by Volker A. Mohnen

The atmosphere functions as a pool and chemical-reaction vessel for a host of substances. Many of the most important ones—oxygen, carbon dioxide and nitrogen and sulfur compounds, for example—are released by the activity of organisms. Often with the help of the water cycle, they pass through the atmosphere and are eventually taken up again into soil, surface water or organic matter. Through technology, human beings have added enormously to the atmospheric burden of some of these substances, with far-reaching consequences for life and the environment. The evidence is clearest in the case of acid rain: precipitation and particles that have been made acidic by air pollution.

The alarm over the increasing acidity of precipitation in Europe and eastern North America was first sounded in the 1960's. Since then the most attention has been focused on acid rain's effects, established and suspected, on lakes and streams, with their populations of aquatic life, and on forests, although the list of concerns is far broader: it includes contamination of groundwater, corrosion of manmade structures and, most recently, deterioration of coastal waters. Twenty years later, how much damage to the ecosystem, lakes and forests in particular, has been confirmed and measured? What has been learned about the processes that produce acid rain and underlie its effects? What does the knowledge imply for efforts to control the emissions—mainly sulfur dioxide from coal- and oil-burning power plants and oxides of nitrogen from motor vehicles and power plants—that cause acid rain?

The study of these questions has grown into a major scientific enterprise. Under the aegis of the National Acid Precipitation Assessment Program (NAPAP), enacted in 1980, many different agencies of the Federal Government sponsor research on the atmospheric processes that produce acid rain, its effects on the ecosystem and options for controlling it. In addition the Electric Power Research Institute, which is funded by the utility industry, supports studies of acid-rain effects and research on technologies for reducing power-plant emissions. The NAPAP will not issue a major report until 1990. Yet much evidence is already in hand—enough to make it clear that acid rain, or more correctly the pollutants that cause it, represents a large-scale interference in the biogeochemical cycles through which living things interact with their environment. Good global housekeeping demands an effort to protect the integrity of these cycles, and economical means of doing so are at hand.

VOLKER A. MOHNEN is professor of atmospheric science at the State University of New York at Albany. He is a graduate of the University of Munich, which awarded him a Ph.D. in 1966, and a past director of SUNY Albany's Atmospheric Sciences Research Center. He has served on several commissions studying atmospheric chemistry and has testified before Congress on the subject of acid rain. In his role as project director of the U.S. Environmental Protection Agency's Mountain Cloud Chemistry Program, Mohnen is currently studying atmospheric processes that may affect the health of forests.

Acid rain is a direct consequence of the atmosphere's self-cleansing nature. The tiny droplets of water that make up clouds continuously capture suspended particles and soluble trace gases. When precipitation coalesces from cloud water, it washes the impurities out of the atmosphere. Not all trace gases can be removed by precipitation, but sulfur dioxide (SO_2) and oxides of nitrogen emitted into the atmosphere are chemically converted into forms that are readily incorporated into cloud droplets: sulfuric and nitric acids.

The processes that convert the gases into acid and wash them from the atmosphere began operating long before human beings started to burn large quantities of fossil fuels; sulfur and nitrogen compounds are also released by natural processes such as volcanism and the activity of soil bacteria. But human economic activity has made the reactions vastly more important. They are triggered by sunlight and depend on the atmosphere's abundant supply of oxygen and water.

The reaction cycle is played out in the troposphere, the lowest 10 or 12 kilometers of the atmosphere. It begins as a photon of sunlight strikes a molecule of ozone (O_3), which may have mixed downward from the ozone layer in the stratosphere or may have been formed in the troposphere by the action of nitrogen- and carbon-containing pollutants. The result is a molecule of oxygen (O_2) and a lone, high-

ly reactive oxygen atom, which then combines with a water molecule (H_2O) to form two hydroxyl radicals (HO). This scarce but active species transforms nitrogen dioxide (NO_2) into nitric acid (HNO_3) and initiates the reactions that transform sulfur dioxide into sulfuric acid (H_2SO_4).

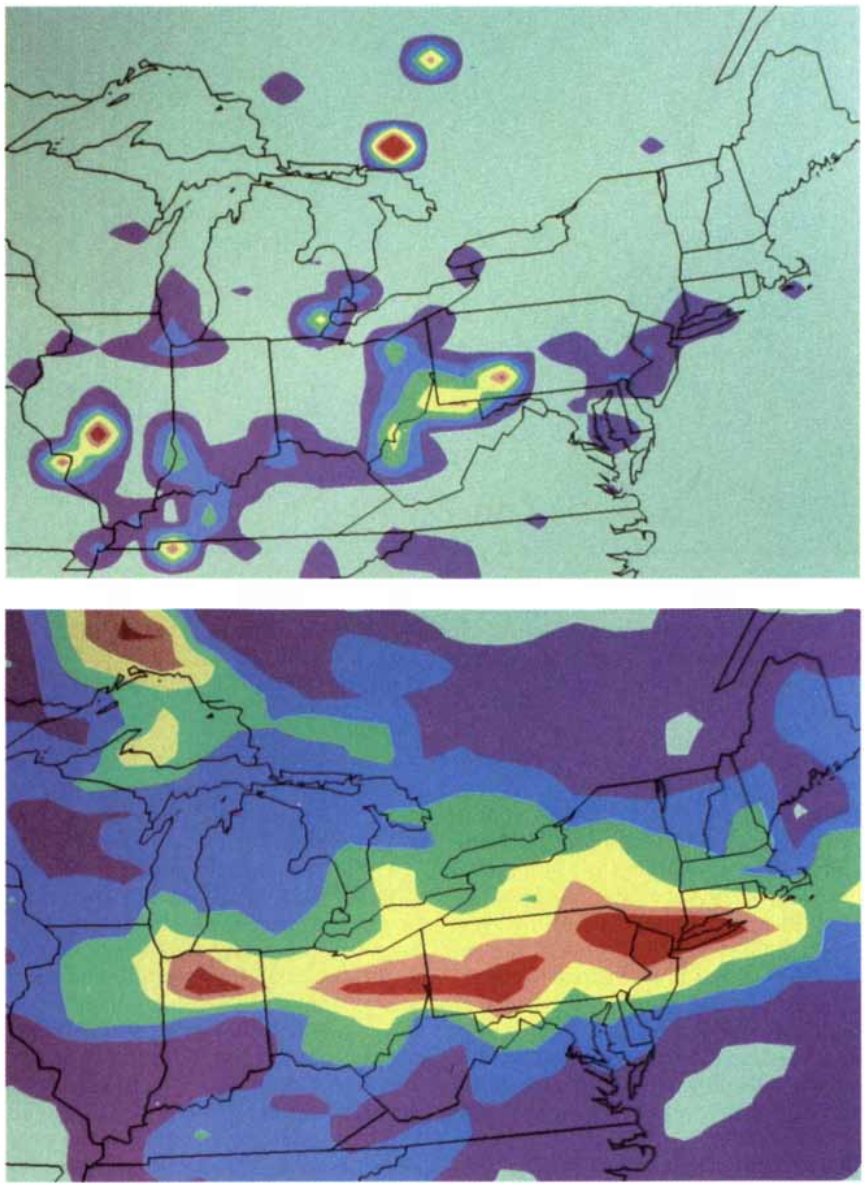
The concentration of the hydroxyl radical in the atmosphere is less than one part per trillion, but it is practically inexhaustible: several of the oxidation processes it triggers end up by regenerating it. For example, one byproduct of the initial oxidation of sulfur dioxide is the hydroperoxyl radical (HO_2), which reacts with nitric oxide (NO) to produce nitrogen dioxide and a new hydroxyl radical. In effect each hydroxyl radical can oxidize thousands of sulfur-containing molecules. As a result only the amount of pollutant in the air determines how much acid is ultimately produced.

The sulfuric and nitric acids formed from gaseous pollutants can easily make their way into clouds. (Some sulfuric acid is also formed directly in cloud droplets, from dissolved sulfur dioxide and hydrogen peroxide.) Nitric acid gas readily dissolves in existing cloud droplets. Sulfuric acid formed through gas-phase reactions condenses to form microscopic droplets, from roughly .1 to two micrometers (millionths of a meter) in diameter, which are one component of the summertime haze in the eastern U.S. Some of these sulfate particles settle to the ground in a process known as dry deposition. (Dry deposition also refers to the capture of sulfur dioxide gas by vegetation.) Most of them, however, are incorporated in clouds. Moisture readily condenses on an existing surface—a condensation nucleus—and sulfate particles are ideal condensation nuclei. They grow into cloud droplets containing dilute sulfuric acid.

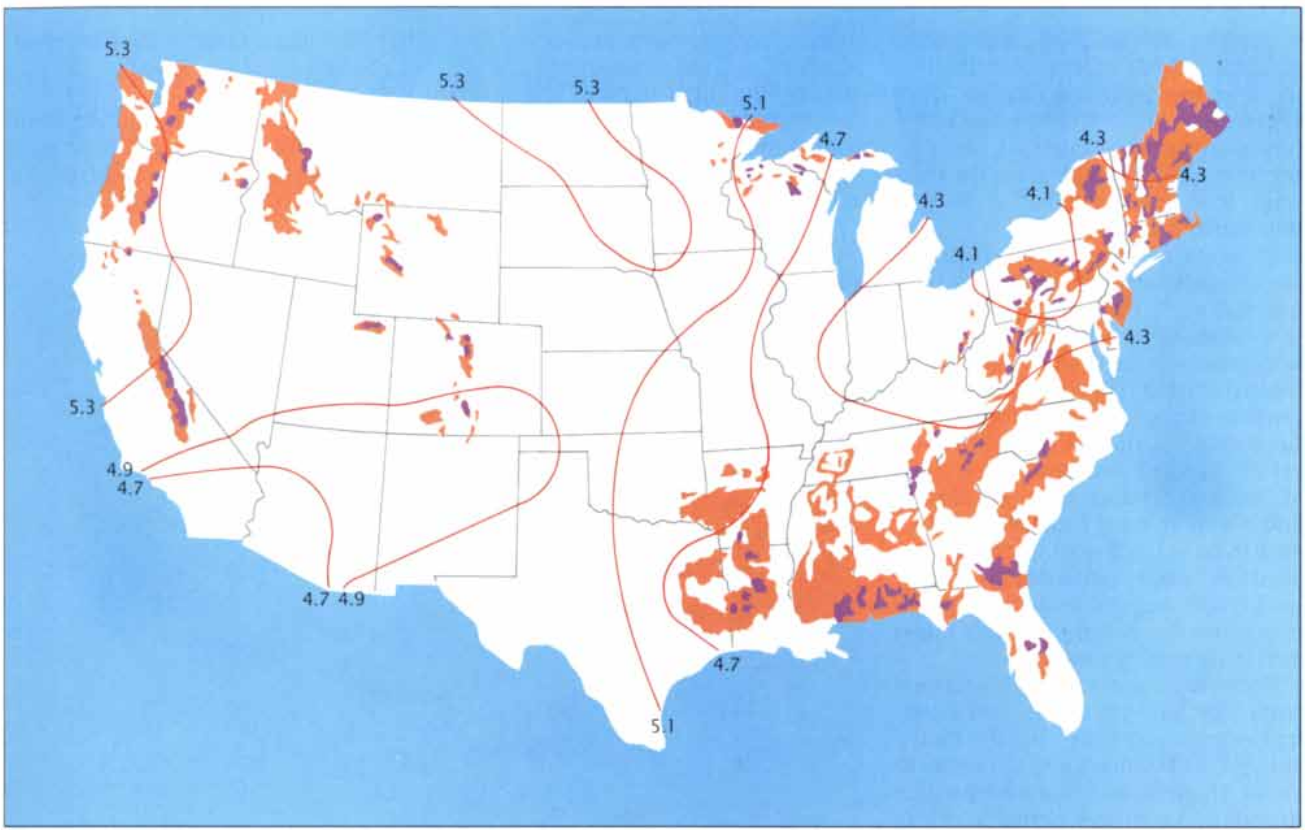
The sulfuric and nitric acids in cloud droplets can give them an extremely low *p*H. Water collected near the base of clouds in the eastern U.S. during the summer typically has a *p*H of about 3.6, but values as low as 2.6 have been recorded. (A *p*H of 7 is neutral; the lower the number, the stronger the acid it represents.) In the greater Los Angeles area the *p*H of fog has fallen as low as 2—about the acidity of lemon juice.

These very high acidities are found only near the base of clouds; the upper reaches are significantly cleaner. Soil and vegetation swathed in acidic clouds, as high-altitude forests can be, are directly exposed to the extremely acidic cloud base. Precipitation particles, however, combine water from much of a cloud's thickness. The resulting dilution of the acid lowers the concentration of sulfur and nitrogen compounds in precipitation by a factor of between three and 30 and the acidity by between one-half and one *p*H unit, to an average in the Northeast of about 4.2.

The acid rain may fall hundreds of miles from the pollution source. Wherever it lands, it undergoes a new round of physical and chemical



SULFURIC ACID DEPOSITION in three days' rain is modeled by computer. The top panel shows the pattern of sulfur emissions (mostly in the form of sulfur dioxide) in the northeastern U.S. and southern Canada, which served as input for the model. The pale background color represents the lowest emission levels and red represents the highest levels. Coal-fired power plants in the Middle West account for most of the U.S. sources; the copper-nickel smelter at Sudbury, Ontario, visible just north of Lake Huron, is the intensest Canadian source. Based on the weather conditions for April 22 through 24, 1981, the computer modeled the transport of the sulfur compounds and other relevant chemicals, their transformation to sulfuric acid and their deposition over the three-day period. The bottom panel shows how the sulfuric acid was deposited in rain; pale color indicates areas that received less than 10 grams of sulfur per hectare (about 2.5 acres) and red indicates areas receiving more than 260 grams per hectare. The computer model, known as the Regional Acid Deposition Model, was developed by Julius Chang of the State University of New York at Albany and his colleagues, with the support of the U.S. Environmental Protection Agency.



THREAT OF ACID RAIN to U.S. lakes and streams is mapped. Brown designates areas in which surface waters tend to have low alkalinity (a low content of ions such as bicarbonate, which can neutralize acid); purple areas have the lowest alkalinity. The contour lines chart the average *p*H of precipitation. Where acidic (low *p*H) precipitation coincides with low surface-water alkalinity, lakes and streams are at risk of becoming acidified. (Alkalinity is not the only factor governing sensitivity to acid rain, however.) The data on water alkalinity were gathered by James M. Omernik of the EPA and his colleagues.

alterations, which can reduce the acidity and change the chemical characteristics of the water that eventually reaches lakes and streams. Alkaline soils, such as soils rich in limestone, can neutralize the acid directly. In the slightly acidic soils typical of the evergreen forests exposed to acid rain in the U.S., Canada and Europe two other processes can blunt the effects of acid deposition. The acid can be immobilized as the soil or vegetation retains sulfate and nitrate ions (from sulfuric and nitric acids respectively). It can also be buffered through a process that is known as cation (positive ion) exchange.

In cation exchange the ions of calcium, magnesium and other metals found in many soils take the place of the acid's hydrogen ions. The source of the metal ions is rock weathering: the dissolution of minerals by precipitation and groundwater containing dissolved carbon dioxide, which releases the positive metal ions into the soil together with anions, or negative ions, of bicarbonate (HCO_3^-). Then, when sulfuric acid is added to the soil, the sulfate (SO_4^{2-}) of the acid can displace the calcium or magnesium ions. As the sulfate solution washes the metal cations from the soil, the hydrogen ions responsible for the acidity are left behind.

The extent to which retention and cation exchange take place in runoff or groundwater depends on the character of the watershed—its geology, vegetation and flow patterns, among other things. Soil processes cannot affect runoff from frozen or fully saturated ground or bare granite bedrock, and so the water that reaches the lake or stream remains about as acidic as the precipitation. Even when the rain does soak in, soil processes may be ineffective. Quartz, for example, is resistant to weathering and lacks the metals needed for cation exchange, and so percolation through quartz sand does little to buffer acid. In watersheds with deep soils capable of retaining large amounts of sulfate or nitrate, however, or soils rich in exchangeable cations, the release of acid to the lake or stream may be forestalled, at least until the retention or buffering capacity is used up.

What happens when acidified runoff or groundwater reaches a lake or a stream? A body of water may contain bicarbonate and other basic ions derived from rock weathering, which can neutralize an influx of acid, preventing the *p*H of the water from falling below a value of about 5. The water's content of such neutralizing ions is known as its acid-neutralizing capacity (ANC), and the value of the ANC provides one measure of a lake's susceptibility to acidification. A lake with a very high ANC is protected against acid rain, at least for the moment; a lake with an ANC of zero may stay healthy if it lies far from acid rain. Otherwise any input of acid will acidify it directly.

An acidified lake is easy to spot. Its ANC is exhausted and its *p*H has fallen to well below 6; its waters are high in sulfate and other ions, such as aluminum, that are mobilized when acid percolates through soil, and it hosts an altered community of aquatic life (or no life at all). Forecasting the acidification of a lake with a low but still positive ANC is another matter. The retention or buffering of acid that is deposited in the watershed may slow the depletion of ANC for the time being. Moreover, a lake's budget of ANC is not fixed. Even as it is depleted by an influx of acid, it may be renewed

by the weathering of minerals in the lake's surroundings. To predict how a lake will respond to a steady input of acid one must know not only its ANC but also how fast its ANC is replenished and how long that rate can be maintained.

These interacting processes in the watershed and the lake, then, determine whether a given lake will acidify, and how fast. They are still not thoroughly understood, and learning enough about a system to predict its behavior is difficult. There is no doubt about the overall trend, however: in areas where the soil is poor in weatherable minerals and acid deposition is heavy, lakes have been acidifying. In 1986 a committee of the National Academy of Sciences compiled measurements of *p*H and alkalinity (a measure of buffering ability similar to ANC) made between the 1920's and the 1940's in several hundred lakes in Wisconsin, New Hampshire and New York and compared the data with recent measurements. In the interim, the committee found, *p*H and alkalinity have on the average increased in the Wisconsin lakes and stayed largely unchanged for those in New Hampshire. In New York, however, and in particular in the Adirondack Mountains, the data for some lakes show a trend of acidification.

The NAS committee got a more complete picture of the trend from microorganisms preserved in lake-bottom sediments. As the *p*H of a lake changes, the assemblage of diatoms and golden-brown algae it hosts changes as well. Species of these minute plants can be distinguished by the form of their skeletons, which makes it possible to reconstruct changes with time in the community of species, and hence in water *p*H. Of the 11 Adirondack lakes for which such data were available, six had increased in acidity since the 1930's, falling to a *p*H of below 5.2; the acidification was fastest during the period ending in 1970. The committee could identify no cause for the *p*H change other than acid rain.

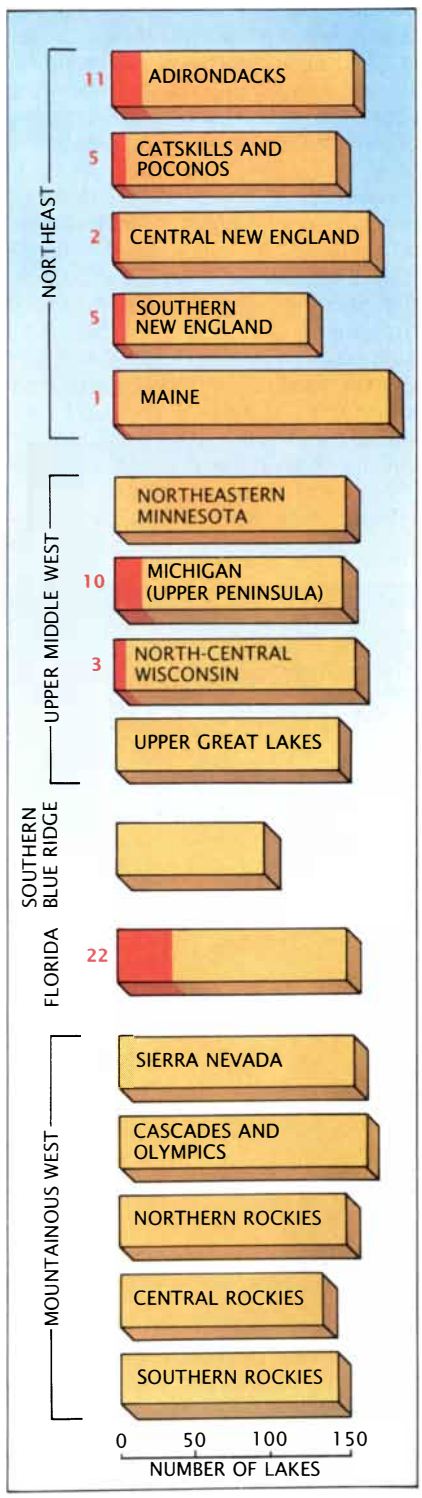
Acidification of lakes in the Adirondacks is a function of the region's highly acidic precipitation (rain collected nearby, in western New York, has an average *p*H of about 4.1, the lowest in the country) and the poor buffering ability of its granite-floored soil and lakes. The recent National Surface Water Survey examined other areas around the country where the ANC of lakes and streams tends to be low, leaving them vulnerable to acid rain. The survey found high percentages of acidic lakes in the Pocono Mountains of eastern Pennsylvania and on Michigan's Upper Peninsula—regions where rain is highly acidic. Acid rain is high on the list of suspected culprits for the relatively large number of acidic lakes found in central and southern New England. Florida showed a strikingly high proportion of acidic lakes, but they are believed to reflect other circumstances, such as organic acids produced by decaying vegetation in swampy regions and fertilizer-rich runoff from agricultural land.

Maine has the lowest percentage of acidic lakes in the Northeast in spite of its poorly buffered soil and waters. Poorly buffered lakes surveyed in the upper Great Lakes region, the southern Blue Ridge Mountains and the mountainous West also were mostly healthy, showing a *p*H of more than 6. What sets those regions apart is their relative freedom from acid rain.

The evidence is not nearly as definitive for the other major environmental effect attributed to acid rain: forest decline. Since 1980 many forests in the eastern U.S. and parts of Europe have suffered a drastic loss of vitality—a loss that could not be linked to any of the familiar causes, such as insects, disease or direct poisoning by a specific air or water pollutant. The most dramatic reports have come from Germany, where scientists, stunned by the extent and speed of the decline, have called it *Waldsterben,* or forest death. Yet statistics for the U.S. are also unnerving.

The decline is most dramatic in high-elevation coniferous forests. For many sites lying above 850 meters in the Adirondacks, the Green Mountains in Vermont and the White Mountains in New Hampshire a comparison of historical records with current surveys shows that more than 50 percent of the red spruce have died in the past 25 years. At lower elevations injury to both softwoods and hardwoods has been documented.

In the forests at high elevations, at least, the dead timber is only the most dramatic evidence of a pervasive loss of tree vigor. Tree-ring records from high-elevation forests in the Northeast show sharply reduced annual growth



ACIDIFIED LAKES are concentrated in the Northeast and the upper Middle West. The graph shows results of the National Surface Water Survey. The segment of each bar in dark color corresponds to the number of sampled lakes whose content of bicarbonate and other acid-neutralizing ions has been depleted; the number is also given as a percentage (*color*). Such lakes usually have a low *p*H and changed aquatic life. Sulfuric and nitric acids from pollutants are thought to account for most such lakes, except in Florida, where they are thought to reflect such factors as organic acids from decaying vegetation and fertilizer runoff.

increments beginning in the early 1960's. The declines occur in stands of many different ages, with different histories of disturbance or disease. What common factor could underlie the growth reductions?

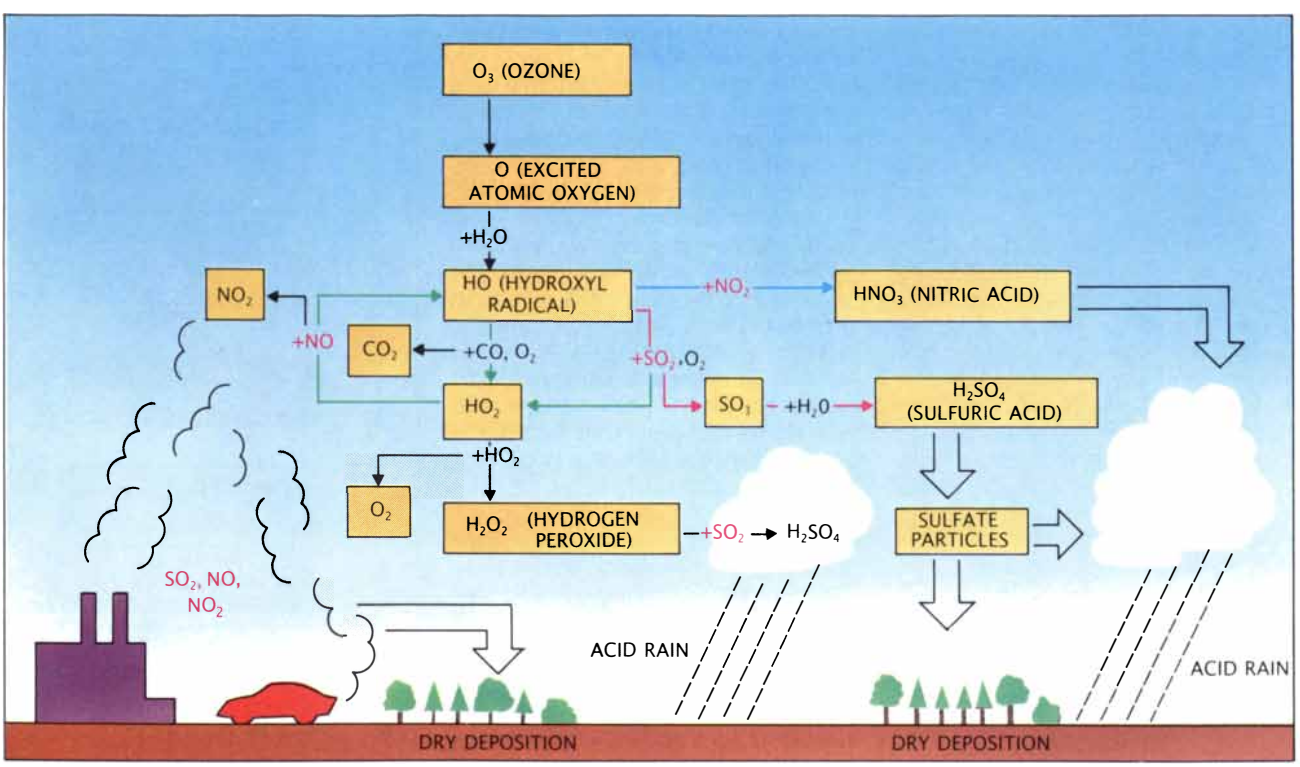
The role of acid rain and other forms of air pollution is under intensive investigation. In spite of the dimensions of the forest damage, however, a firm causal link has proved to be elusive. One can get some idea of the difficulties by contrasting the recent forest decline with clear-cut cases of fumigation: forest poisoning by air pollutants. Smelters and chemical plants that emit sulfur dioxide, oxides of nitrogen or fluoride compounds are often girdled by dead timber. In such cases there is a clear correlation between tree damage, a specific pollution source and a threshold concentration of the pollutant. The forests that are now dying, in contrast, are far from any source and are exposed to pollutants in concentrations well below the levels previously reported to injure trees. If air pollution, and specifically acid rain, plays a part in forest decline, it probably does so less as a lethal agent than as a stress.

Many stresses, both biotic and abiotic, combine to affect the vigor of a forest. The trees' genetic endowment or age can be a source of stress: a stand may be genetically weak or senescent. Other stresses may take such forms as diseases, insects, parasitic fungi and seed plants, a shortage of light, water or essential nutrients and sporadic injury from events such as floods, high winds and ice storms. Stresses easily withstood in isolation can combine with debilitating or fatal effects. A fatal sequence of stresses may begin with a "predisposing" stress, such as a shortage of nutrients. The tree may then be seriously weakened by an "inciting" stress, such as a severe winter. It is then defenseless against a final, "contributing" stress—the actual cause of death—such as disease or insect attack.

Acid and other pollutants could add to the high level of abiotic stresses, including thin soil, low temperatures and desiccating winds, present in a high-elevation forest. That is, the pollutants might handicap the trees with one more predisposing stress as they face subsequent stresses. But what is the nature of the added stress?

Investigators, most of them in Europe, have put forward a number of hypothetical mechanisms, many of which would ultimately lead to nutrient deficiency in the tree. Several mechanisms would be played out in the soil. The aluminum released from soil minerals by acid might compete with calcium for binding sites on fine roots, reducing a tree's supply of calcium and slowing its growth. Alternatively, the soil itself might lose nutrients when vital elements such as calcium, magnesium and potassium are leached away by acid rain. The death of soil microorganisms is another possible source of nutrient stress. Low soil *p*H and high concentrations of aluminum can reduce populations of the bacteria that break down and release nutrients locked in decaying organic material. In addition, high levels of nitrate from nitric acid deposition can injure the mycorrhizae, symbiotic fungi that live on the roots of conifers and help the trees to ward off disease and extract water and nutrients.

In other scenarios the pollutants would work their effects aboveground. Acid rain or, more likely, acidic cloud droplets intercepted by the needles of a conifer could leach out nutrients—magnesium, calcium and potassium in particular—faster than the tree's roots



ATMOSPHERIC CHEMISTRY generates sulfuric and nitric acids from sulfur dioxide and oxides of nitrogen given off by industry and vehicles. The hydroxyl radical, formed when a molecule of ozone breaks apart and releases an oxygen atom that can react with water, is the major actor. It converts nitrogen dioxide (NO_2) into nitric acid (*blue*) and initiates the conversion of gaseous sulfur dioxide (SO_2) into sulfuric acid (*red*). (A different reaction sequence forms sulfuric acid from sulfur dioxide and hydrogen peroxide dissolved in cloud water.) The hydroxyl radical is regenerated by reactions (*green*) involving nitric oxide (NO), and the acids come to the earth as dry particles and in rain and other forms of precipitation.

could replace them. An additional pollutant, ozone, might worsen nutrient leaching by degrading the water-resistant waxy coating of the needles. Still another hypothesis holds that ozone alone might lead to nutrient stress because it can damage chlorophyll, thereby impeding photosynthesis.

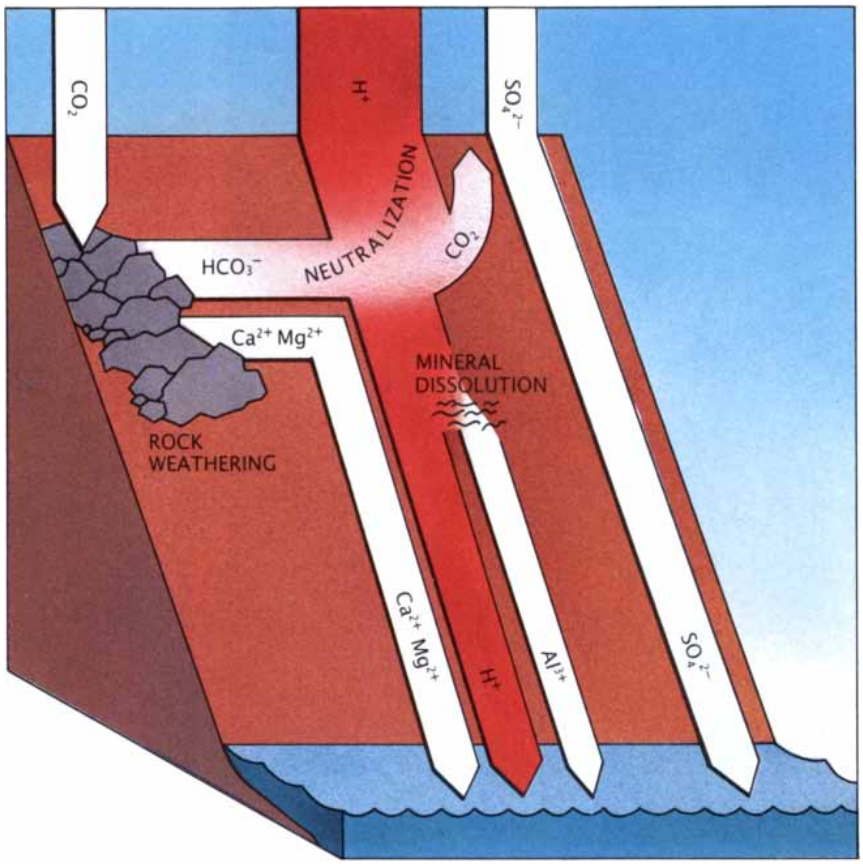
Finally, acid rain might augment the stress of low winter temperatures. In the fall a conifer ordinarily prepares for the freezing temperatures of winter by withdrawing water from its needles, a process known as cold hardening. The initiating signal for cold hardening ordinarily comes from the roots, in the form of a decreased supply of the nitrogen-bearing nutrients that are produced by soil microorganisms. As acid soaks into the needles, however, the nitrogen compounds it contains might in effect fertilize the tree. They might override the signal from the roots, delaying cold hardening and leaving the tree vulnerable to damage from ice formation in needle tissue. Ozone too might reduce a tree's resistance to freezing by damaging cell membranes in the foliage.

Laboratory tests are now under way to see which of these mechanisms (if any) might operate under the conditions of pollutant exposure in the afflicted forests. But only field studies, in the forests themselves, can show that a given mechanism is actually at work. The task is challenging: one is trying to track down what may be a relatively small increment of stress, superimposed on a complex set of natural stresses. That background of stresses may vary from stand to stand and even from tree to tree.

Whiteface Mountain in the Adirondacks provides a case in point. It displays some of the most dramatic forest decline in the U.S., but because of the dominance of several natural stresses only tentative conclusions about the role of pollutants can be drawn. The direct cause of forest decline, inferred from foresters' records and temperature data, seems to have been severe, repeated damage by desiccation or freezing during the winters in the early 1960's. Ozone may well have made the trees more vulnerable to frost damage: shielding tree limbs from the ambient ozone leads to changes in biochemistry that suggest ozone can indeed weaken the tree by attacking cell membranes in the foliage. The role of acid rain and acidic clouds has not yet been fully investigated, but it is conceivable that they also acted as a predisposing stress in some way.

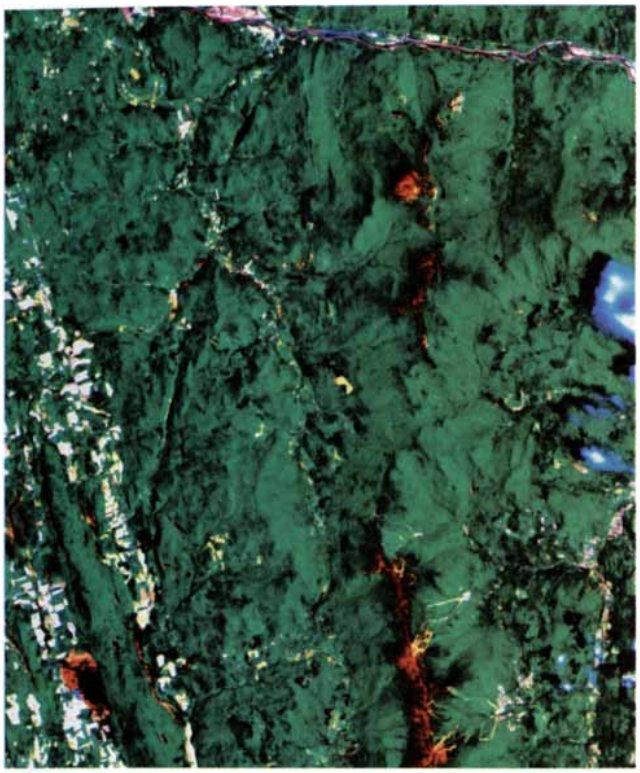
Even though uncertainties surround acid rain's role in forest decline, its effects in the soil and water alone leave no question about the need to reduce the ambient burden of sulfur and nitrogen compounds and thereby lower the acidity of precipitation. Some progress has already been made. In the Northeast the sulfate content of rain and the concentration of airborne sulfur compounds have decreased in the past 15 years; the decreases reflect the pollution-control measures mandated by the Clean Air Act, enacted in 1975, and additional emission laws passed by individual states. The rate at which lakes in the Northeast are acidifying seems to have slowed as well. To actually reverse the trend, however, acid deposition will have to be reduced much further, and many policymakers and scientists are now asking: How quickly? By how much?

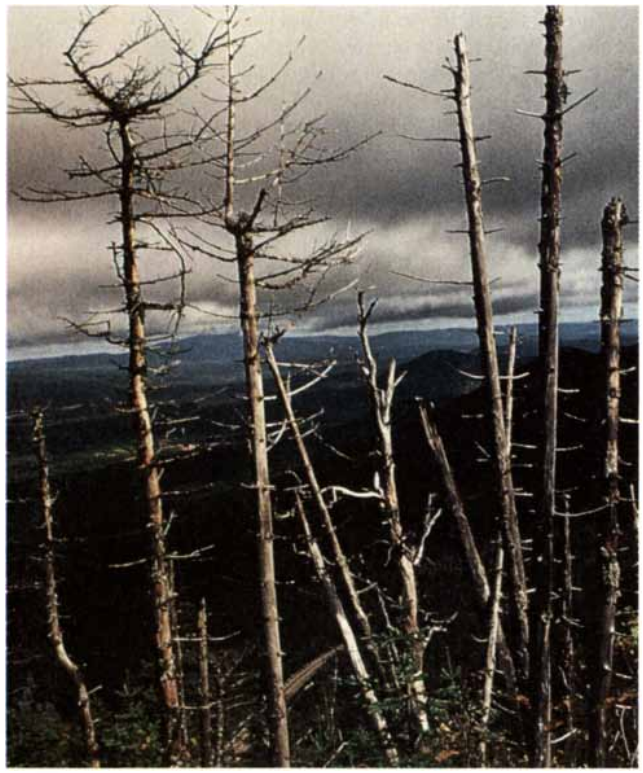
For precise answers to those questions we need to know how long soil processes can continue to buffer or retain acid in the threatened regions and how fast lakes can renew their acid-neutralizing capacity. We also need to understand the relation between acid rain and forest decline. Some answers should be forthcoming in the 1990 NAPAP report. Certain scientists have already speculated, however, that to protect lakes and streams in sensitive areas such as the Adirondacks it will be necessary to reduce acid deposition to less than 50 percent of its current level.



WATERSHED PROCESSES can alter the chemistry of acidic rainwater before it reaches a lake or stream. The illustration shows several processes that can act on sulfuric acid percolating through a hillside. So-called cation exchange can take place if the soil is rich in cations (positive ions) such as calcium and magnesium (Ca^{2+} and Mg^{2+}). Such ions are released from certain rocks by the weathering action of groundwater or precipitation containing dissolved carbon dioxide, a process that also generates bicarbonate ions (HCO_3^-). Some of the acid's hydrogen cations (*red*) then displace calcium and magnesium and are themselves retained in the soil, where bicarbonate ions can neutralize them. These processes reduce the hydrogen-ion concentration—the acidity—of water reaching the lake or stream. The acid can also dissolve clay minerals in the soil and release aluminum, which can harm plants and aquatic life.

Where and by how much will emissions have to be reduced to achieve such reductions in deposition? Guidance will come from two massive computer models of acid production, transport and deposition that are now being tested: the Regional





DEAD OR DYING FOREST on high slopes of the Green Mountains in Vermont is apparent as areas of red in a false-color satellite image made at infrared wavelengths that are sensitive to chlorophyll (*left*). A photograph made on the ground at Whiteface Mountain in New York shows dead spruce (*right*). Many investigators think acid rain, perhaps in combination with other pollutants, has caused the rapid decline of some alpine forests in the eastern U.S., although a causal link has not been established. The satellite image is from the Landsat *Thematic Mapper* and was provided by James E. Vogelmann of the University of New Hampshire; the photograph of the spruce was provided by Ann Carey of the U.S. Forest Service.

Acid Deposition Model (RADM), supported by the U.S. Environmental Protection Agency, and the Acid Deposition and Oxidant Model (ADOM), supported by agencies of the Canadian and West German governments. The models take into account all the atmospheric chemistry and meteorological processes known to act on molecules containing sulfur, nitrogen and carbon. (Carbon-containing molecules are included because of their role in producing the oxidants that convert sulfur and nitrogen emissions into acids.)

Given a set of source locations, emission levels and atmospheric conditions, these models can forecast weather and atmospheric chemistry in order to predict, with a geographic resolution of better than 50 miles square, the amount of acid deposited across an entire region in the course of up to four days. By averaging results calculated for a variety of atmospheric conditions, the models can also predict the long-term pattern of deposition for a given emission pattern, which should make them invaluable for designing a strategy of emission reductions.

How might the cuts be made? The most direct way of controlling the pollutants that cause acid rain would be to burn less fossil fuel for transportation and energy generation. Expanded mass transit and fuel-efficient cars can reduce oil consumption in the transportation sector, but energy generation is less tractable. In spite of worthy strategies for conserving energy, consumption is likely to increase in the long run, and current alternatives to fossil-fueled power plants do not look promising. Hydroelectric power is limited by a scarcity of appropriate sites, and nuclear power is beset by economic problems and a crisis of public confidence in its safety.

The key to controlling acid rain, then, must be the reduction of emissions from fossil-fueled power plants, coal-burning plants in particular. The approach that has already led to reductions in sulfur emissions in the U.S., West Germany and Japan combines the use of coal that is naturally low in sulfur, or has been washed to remove sulfur and other contaminants, with flue-gas desulfurization (FGD). In FGD wet limestone is sprayed into the plant's hot exhaust gases, where it scavenges as much as 90 percent of the sulfur dioxide. The sulfur-containing waste can be difficult to dispose of, however, and FGD reduces the efficiency of a power plant, causing it to consume several percent more coal for a given output. Furthermore, the process does nothing to reduce nitrogen oxide emissions.

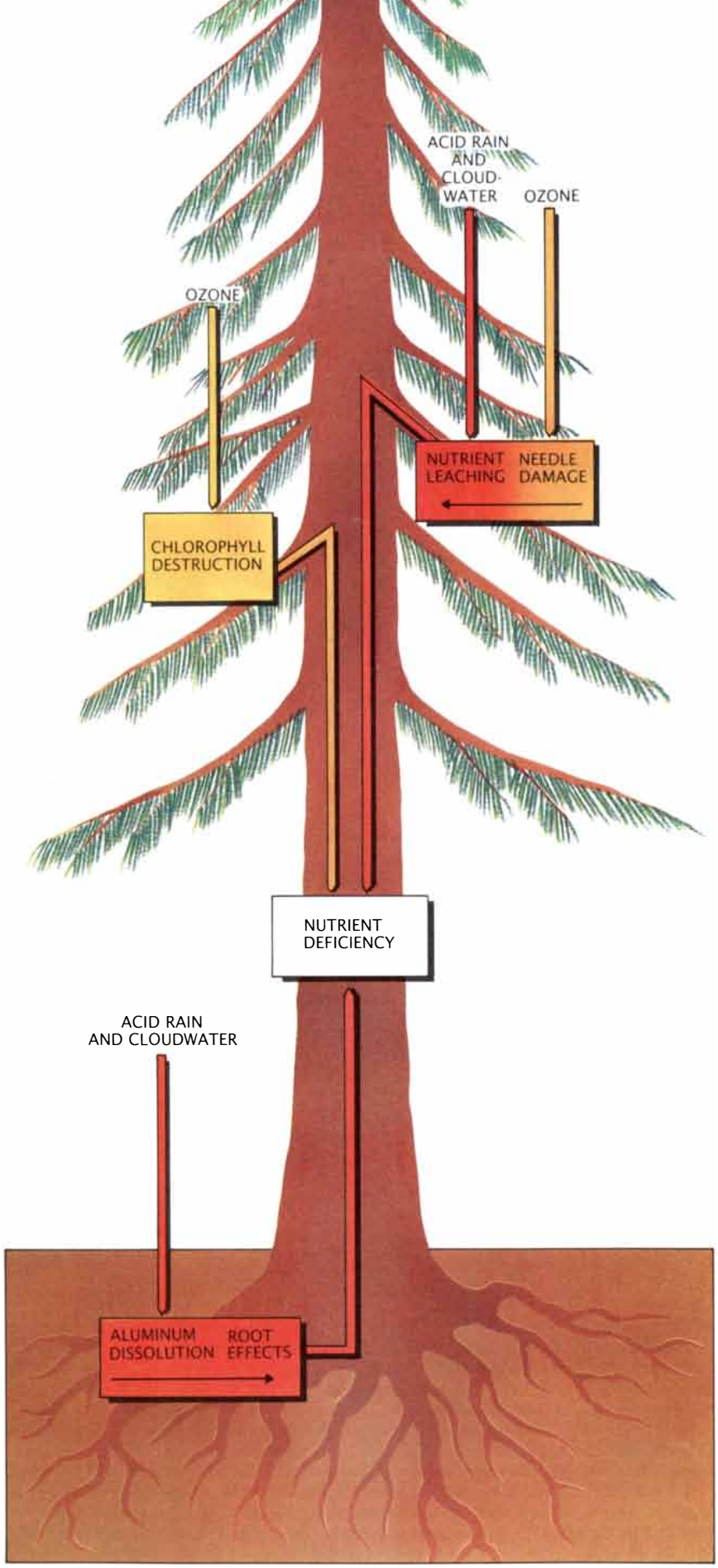
The new power-plant technologies developed jointly by the Government and industry under the Clean Coal Demonstration Program, enacted in 1984, offer a more comprehensive solution. Three clean-coal technologies are already being demonstrated in full-size plants [see "Coal-fired Power Plants for the Future," by Richard E. Balzhiser and Kurt E. Yeager; SCIENTIFIC AMERICAN, September, 1987]. In the system known as atmospheric fluidized-bed combustion, a turbulent bed of pulverized coal and limestone is suspended by an upward blast of air; the combustion region is threaded with boiler tubes, which supply steam to the plant's turbines. The turbulent mixing of the coal and the air allows combustion to take place at a lower and more even temperature than it

does in a conventional boiler, which reduces the formation of oxides of nitrogen. Meanwhile the limestone efficiently captures the sulfur dioxide. In a related technology known as pressurized fluidized-bed combustion the coal is burned in compressed air, which improves the plant's efficiency as well.

In the third technology, gasification/combined-cycle, coal is reacted with steam and air at high temperatures to produce a gas consisting mainly of hydrogen and carbon monoxide. The gas can then be burned, spinning a turbine; waste heat in the gas turbine's exhaust serves for generating steam, which drives a steam turbine to yield additional electricity. A gasification/combined-cycle plant operates much more efficiently than a conventional plant and gives off considerably less sulfur dioxide and nitrogen oxides.

Retrofitting existing plants with FGD offers the fastest way to reduce power-plant emissions. Almost half of the coal-fired plants in the U.S. were built before 1975 and have no controls for sulfur and nitrogen pollutants. Concentrated in the eastern half of the U.S., they account for most of the country's sulfur emissions. Adding conventional FGD to the plants could cut total emissions of sulfur dioxide from all power plants to less than half their present level, and the reduction could be accomplished within 15 years. Emissions of nitrogen oxides would not be affected, however. In addition, utilities object to the expense of installing and operating FGD equipment and the loss of plant efficiency it would cause.

Clean-coal technologies present an attractive alternative. Any effort to control acid rain must be focused on the aging plants, many of which will soon become candidates for retirement or refurbishment. Gradually re-



ACID RAIN AND OZONE together could lead to nutrient deficiency in a coniferous tree, according to a currently favored scenario for their possible role in forest decline. The ozone might act both by destroying chlorophyll (vital to photosynthesis) and by degrading the waxy coating of the needles. Acid rain or cloud water could then soak into needle tissue more readily and leach out nutrients. In the ground the acid might compound the nutrient deficiency by mobilizing aluminum, which could displace calcium from its binding sites on the tree's fine roots. Stressed by lack of nutrients, the tree would be vulnerable to destruction by insects, disease or some other process.

placing them with new conventional plants equipped with FGD would yield only modest reductions in emissions, and the cost of designing, building and getting regulatory approval for the new plants would be staggering. Instead most of the old plants—the 410 generating stations built between 1955 and 1975—could be "repowered": refurbished with a new combustion section incorporating one of the clean-coal technologies.

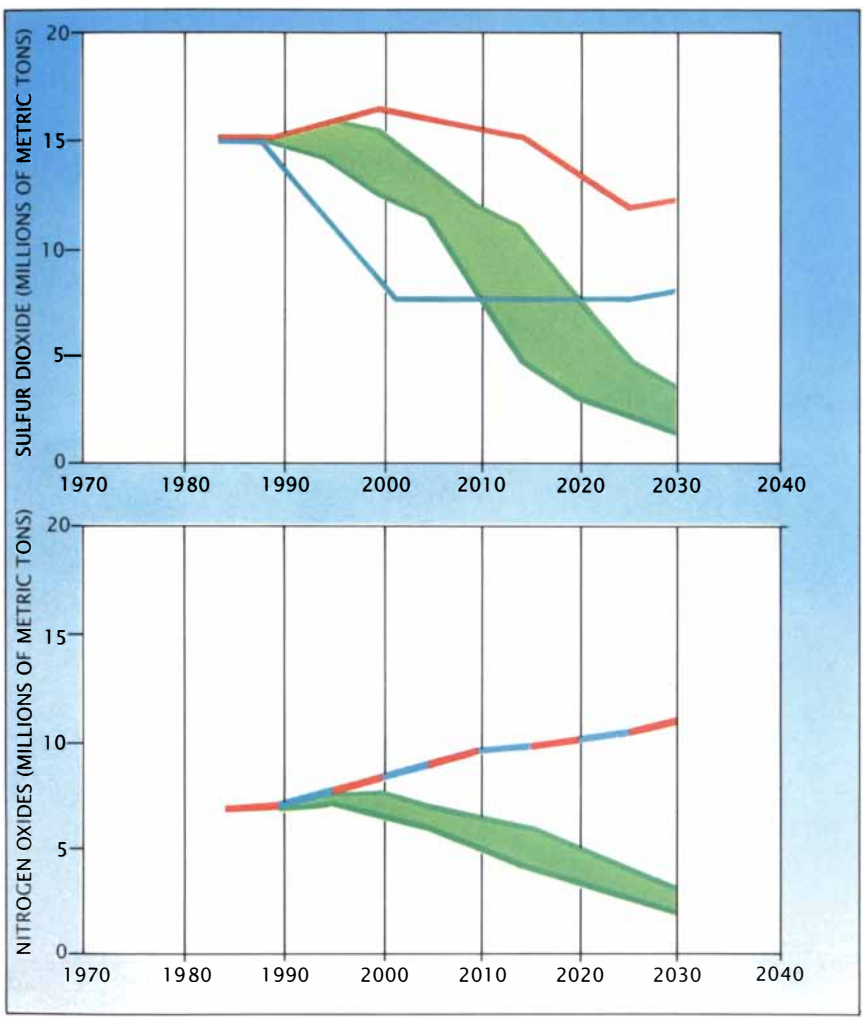
A repowered plant could preserve much of its existing equipment for handling coal and ash and most of its steam-cycle and electricity-generating hardware. The repowering of an existing plant would thus be quick and cheap compared with building a new one. The approach has an additional attraction for the utility industry: the new hardware could be added to a plant in modules, which would enable utilities to adjust generating capacity to the demand for power.

The repowering of aging plants promises ultimately the greatest emission reductions, affecting the full range of pollutants implicated in acid rain. The strategy could cut sulfur dioxide emissions by more than 80 percent and nitrogen oxide emissions by more than 50 percent; the emission of fewer nitrogen compounds would in turn reduce the formation of ozone in the troposphere.

We now know that the term acid rain covers a host of phenomena. Oxides of nitrogen, for example, affect the chemical cycle that converts sulfur dioxide into sulfuric acid, and the ozone they help to produce may work in concert with acid rain to destroy forests. The nitrate ion as well as the acidity that accompanies it may affect the ecosystem, not just on land but also, it now appears, in coastal waters. The emission reductions that are promised by repowering could lessen all these effects.

The drawback is that the reductions, dramatic as they are, would be slow in coming. The recent declines in the sulfate content of air and precipitation in the Northeast and the slowing of lake acidification suggest that some breathing space remains. The nation can probably forgo the short-term solution of retrofitting pollution controls on existing plants in favor of the gradual but more comprehensive and economical approach of repowering. Yet Government intervention, in the form of a timetable, may still be needed to speed the pace. If utilities simply repower plants as the need arises—as the plants reach an age of 50 years or so—the process would not be completed until well into the next century.

Technology has leapfrogged science and presented us with an option for addressing the problem of acid rain that is likely to be attractive whatever the resolution of the remaining scientific uncertainties. The urgent need to reduce human interference in the complex chemistry of the biosphere is already painfully clear.



EXPECTED REDUCTIONS in annual emissions of sulfur dioxide (*top*) and oxides of nitrogen (*bottom*) from power plants vary depending on the choice of technology. Replacing plants as they reached 50 years of age with new ones incorporating flue-gas desulfurization (FGD) would reduce only sulfur dioxide, and the reductions would come slowly (*red*). Retrofitting all existing power plants with FGD on a 15-year schedule would yield much sharper reductions, again in sulfur dioxide alone (*blue*). In both cases nitrogen oxide emissions would continue to climb as additional plants were built to satisfy growing demand for power. Refurbishing plants built between 1955 and 1975 with "clean coal" technologies such as fluidized-bed combustion (a strategy known as repowering) would eventually lead to the largest cuts in both kinds of pollutants. The green bands illustrate the range of reductions expected if the plants were refurbished when they reached an age of between 40 and 50 years.

FURTHER READING

ACID RAIN. Gene E. Likens, Richard F. Wright, James F. Galloway and Thomas J. Butler in *Scientific American,* Vol. 241, No. 4, pages 43–51; October, 1979.

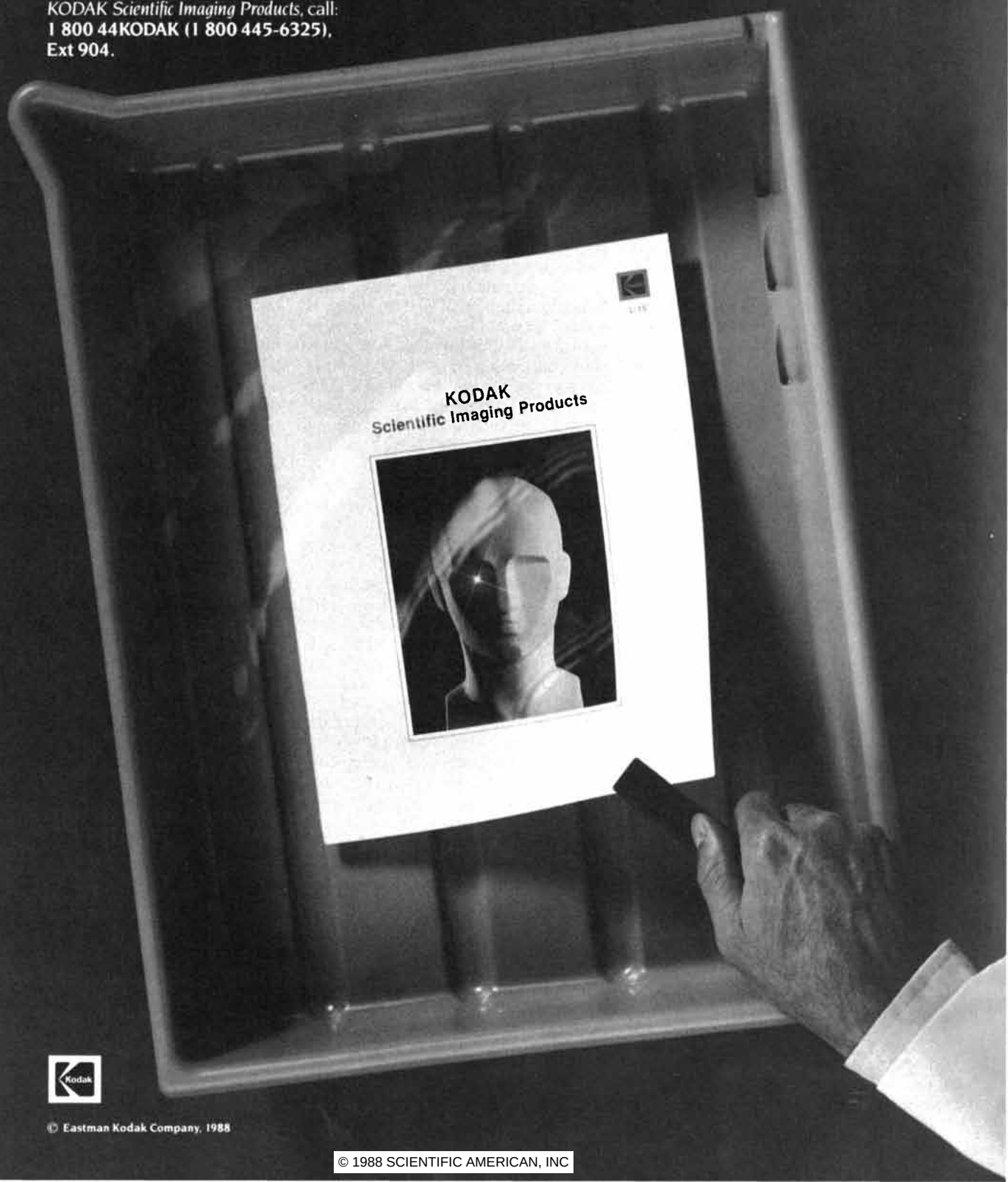
SPECIAL ISSUE ON THE EXPERIMENTAL LAKES AREA. In *Canadian Journal of Fisheries and Aquatic Science/Journal canadien des sciences halieutiques et aquatiques,* Vol. 21, No. 37, pages 313–558; March, 1980.

ACID PRECIPITATION IN HISTORICAL PERSPECTIVE. Ellis B. Cowling in *Environmental Science and Technology,* Vol. 16, No. 2, pages 110a–123a; February, 1982.

AIRBORNE CHEMICALS AND FOREST HEALTH. James N. Woodman and Ellis B. Cowling in *Environmental Science and Technology,* Vol. 21, No. 2, pages 120–126; February, 1987.

Now Appearing For The Professional.

Now. The most complete selection of Kodak scientific imaging products ever included under one cover. Plates films, papers, filters, chemicals, and more—arranged according to applications. And enlightened with our recommendations. For your free copy of *KODAK Scientific Imaging Products*, call: **1 800 44KODAK (1 800 445-6325), Ext 904.**



The High Fidelity of DNA Duplication

Generation after generation, through countless cell divisions, the genetic heritage of living things is scrupulously preserved in DNA. Why are so few mistakes made when the DNA is copied?

by Miroslav Radman and Robert Wagner

A tale of uncertain origin has it that Caesar, intending to grant amnesty to one of his army officers, issued the order "Execute not, liberate." His message, however, was passed along with one small mistake: the comma was misplaced. The message the officer's guards received read "Execute, not liberate" and the unfortunate man lost his life.

All of life depends on the accurate transmission of information. As genetic messages are passed along through generations of dividing cells, even small mistakes can be life-threatening. In human beings the substitution of a single "letter" in the genetic message is responsible for such lethal hereditary diseases as sickle-cell anemia and thalassemia. Several common cancers are also associated with a single-letter change.

To be sure, not every mistake in the transmission of genetic messages is catastrophic; some mistakes make no difference at all, and some can even be beneficial. But for the most part the message must be accurately preserved and transmitted.

For organisms as complex as human beings, attaining sufficient accuracy is a monumental feat. The set of genetic instructions for humans is roughly three billion letters long. If mistakes were as rare as one in a million, 3,000 mistakes would be made during each duplication of the human genome. Since the genome replicates about a million billion times in the course of building a human being from a single fertilized egg, it is unlikely that the human organism could tolerate such a high rate of error. In fact, the actual rate of mistakes is more like one in 10 billion. How do cells achieve such high fidelity?

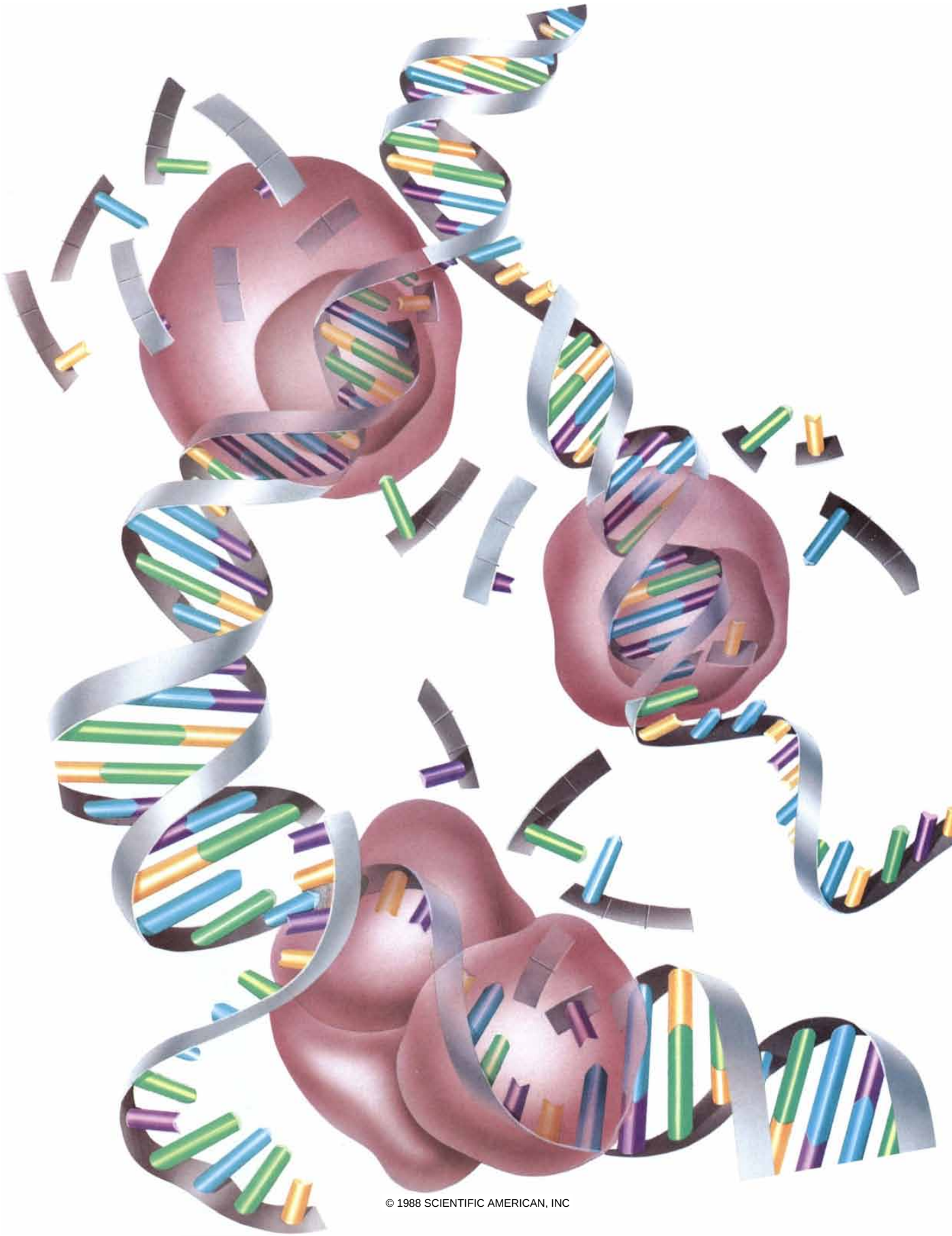
Evidence drawn from many different disciplines has revealed that discrete enzymatic processes cooperate to ensure the hi-fi replication of DNA. It may be that nature has evolved several complementary mechanisms because a single very thorough system would be too time- and energy-consuming. The occasional errors that slip through the hi-fi mechanisms are also significant, since mutations are an important source of the genetic variation that is necessary for adaptive flexibility.

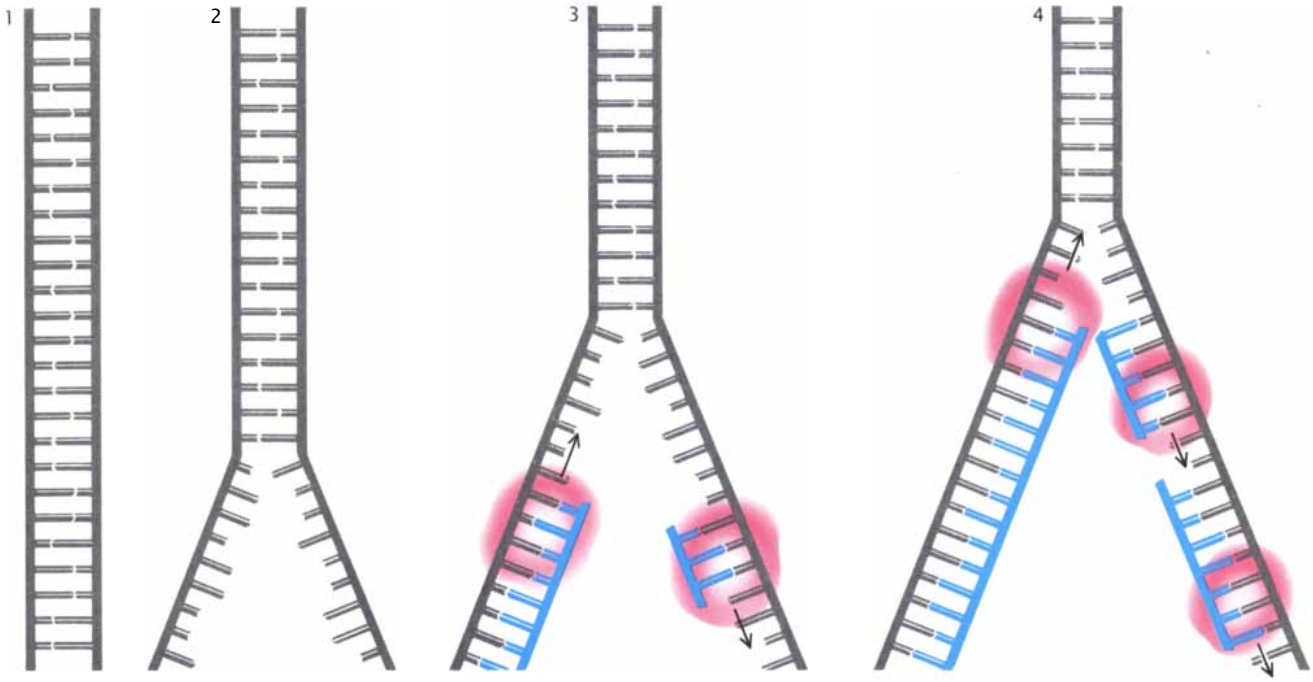
In the cells of all living organisms the genetic message is contained in double-stranded DNA. DNA's structure is marvelously suited to maintaining the integrity of the genetic message. The two strands are complementary, which is to say they carry the same genetic information, in the sense that positive and negative strips of movie film portray the same scene. Like the strips of film, one strand of DNA can be used to reconstruct the other. If one strand is damaged, it can be repaired by removing the damaged region and using the undamaged strand as a template for synthesizing a new strand. Indeed, DNA is routinely replicated by a similar process: the two parent strands are separated at a "replication fork," and each becomes a template for building a new strand [*see illustration on page 42*].

The biochemical "letters" that encode information in DNA are four nucleotides, which are distinguished by the bases they contain. The bases are adenine, guanine, thymine and cytosine, commonly designated *A, G, T* and *C*. The order in which nucleotides occur determines the "meaning" of the genetic message. The bases on one strand pair with the bases on the other strand, linking the two strands like rungs on a ladder. The pairing is not random: adenine must pair with thymine, and guanine must pair with cytosine. Hence the complementarity of

MIROSLAV RADMAN and ROBERT WAGNER have collaborated on numerous research projects and papers, although their careers have taken radically different paths. Radman is a research director at the National Center for Scientific Research (CNRS) in Paris. He was born in Yugoslavia and received a Ph.D. from the Free University of Brussels in 1969. After 10 years on the faculty at the Free University, he joined the CNRS in 1983. Radman and Wagner met in 1972 at Harvard University, where Radman was a research fellow and Wagner was a graduate student in Matthew S. Meselson's laboratory. Wagner left Harvard to try farming soon after getting his Ph.D. in 1976. Four years later he moved to Brussels to resume his research with Radman. In 1981 he returned to the U.S. and bought a 375-acre farm near the Canadian border in upstate New York, where he now raises purebred Suffolk sheep. Wagner still spends several weeks of the year working with Radman at his laboratory in Paris.

REPLICATING DNA is subjected to three different "quality control" mechanisms, which are represented by the three enzyme complexes (*brown*) in this schematized drawing. During replication the two strands of the parental double helix are separated so that they can serve as templates for the synthesis of two new DNA strands. The nucleotides (*colored pegs*) of the new strands should be complementary to the nucleotides of the old ones: blue pairs with purple, and green pairs with yellow. The enzyme complex at the upper left is screening nucleotides for incorporation into the elongating strand. The complex at the right is excising an incorrect nucleotide that has just been added to the strand. The complex at the lower left is repairing a gap where a fragment with a mismatched nucleotide is being cut out.





REPLICATION FORK divides the double-stranded DNA molecule (*1*) like a zipper (*2*). Synthesizing enzymes called DNA polymerases (*red*) attach to the bared parental strands, and synthesis of the two nascent strands (*blue*) proceeds in opposite directions (*3*). The polymerase on one side tags along after the advancing fork. On the other side new polymerases must keep binding close to the fork to synthesize the strand between the fork and the site at which the last polymerase bound (*4*).

base pairs is the basis for the complementarity of DNA strands.

Errors arising in the course of DNA synthesis can result in noncomplementary base pairs, or mismatches. Other kinds of errors can be introduced by environmental influences. Repair of environmentally inflicted damage to DNA (from chemicals, radiation and so on) has been extensively studied and reviewed [see "Inducible Repair of DNA," by Paul Howard-Flanders; SCIENTIFIC AMERICAN, November, 1981]. This article will concern itself only with those mistakes that arise during DNA replication. When DNA is synthesized in the absence of enzymes, such errors happen about once in every 100 bases. The enzymatic systems discussed here make synthesis 100 million times more accurate than nonenzymatic synthesis.

Three enzymatic processes are responsible for the high fidelity of DNA replication. The first process is involved in selecting which of the four nucleotides is added to the nascent strand. The second process involves "proofreading" the most recently added nucleotide and expelling it if it is noncomplementary. The third process takes place after synthesis and involves correcting errors that escaped the first two "editors." Because nucleotide selection and proofreading act in concert with the DNA replication machinery, they are known as error-avoidance mechanisms. The mechanism that operates after synthesis is an error-correction mechanism; it is called mismatch repair.

The accuracy of replication is due primarily to the effectiveness of nucleotide selection. The selection is mediated by the same enzyme that carries out the polymerization of nucleotides. The enzyme, called DNA polymerase, moves along the DNA template and synthesizes the complementary strand from the cellular pool of nucleotides. The free nucleotides are in the form of triphosphates; that is, they carry a string of three phosphate groups. The nucleotides must be cleaved to monophosphates before they can be added to the new strand. DNA polymerase takes up a nucleotide triphosphate, cleaves it to a monophosphate and adds the latter to the end of the nascent strand.

Nucleotide selection depends on the energetic relations between competing reactions; in other words, it is possible to insert any base opposite any other base, but the correct pairing is the most energetically favorable. Evidence suggests that selection acts both on triphosphate binding and on monophosphate incorporation [*see illustration on opposite page*]. In the case of triphosphate binding, for instance, results from experiments by Nancy Nossal of the National Institute of Arthritis, Diabetes, and Digestive and Kidney Diseases and by David Korn of the Stanford University School of Medicine have shown that the complex of polymerase, template DNA and nucleotide triphosphate is stablest when the triphosphate is complementary to the template nucleotide.

Selection at the level of monophosphate incorporation is probably more rigorous than the selection that acts on triphosphate binding. A model based on work in our laboratory (which was then at the Free University of Brussels) postulates that most of the nucleotides the polymerase binds make it through the triphosphate screen; that is, they get cleaved to monophosphates and aligned with the template strand. If a nucleotide is indeed complementary, it fits well with the template base and the addition is stabilized. If the nucleotide is noncomplementary, it does not fit as well, the reaction is reversed and the nucleotide is restored to its triphosphate form. An apt metaphorical role for the polymerase would be that of a blind cook, who grabs ingredients at random, tastes each one and decides whether to add it to the soup or put it back on the shelf.

At the level of nucleotide selection, noncomplementary nucleotides are incorporated at the rate of about one in 100,000. An error that slips through this process encounters the second

mechanism of error avoidance: proofreading. Proofreading is carried out by an enzymatic activity that is either part of or associated with the DNA polymerase. This activity was nicknamed "proofreading exonuclease" by Stanford investigators Douglas L. Brutlag and Arthur Kornberg, who discovered it in the early 1970's. The exonuclease is capable of removing both complementary and noncomplementary nucleotides from the terminal of the nascent chain. However, as a rule it only gets the opportunity to act when a nucleotide is noncomplementary. The presence of the mismatched nucleotide greatly inhibits the addition of the next nucleotide, and the pause in the polymerization process gives the exonuclease time to remove the noncomplementary nucleotide. The polymerase then tries again to find a complementary nucleotide for the terminal position.

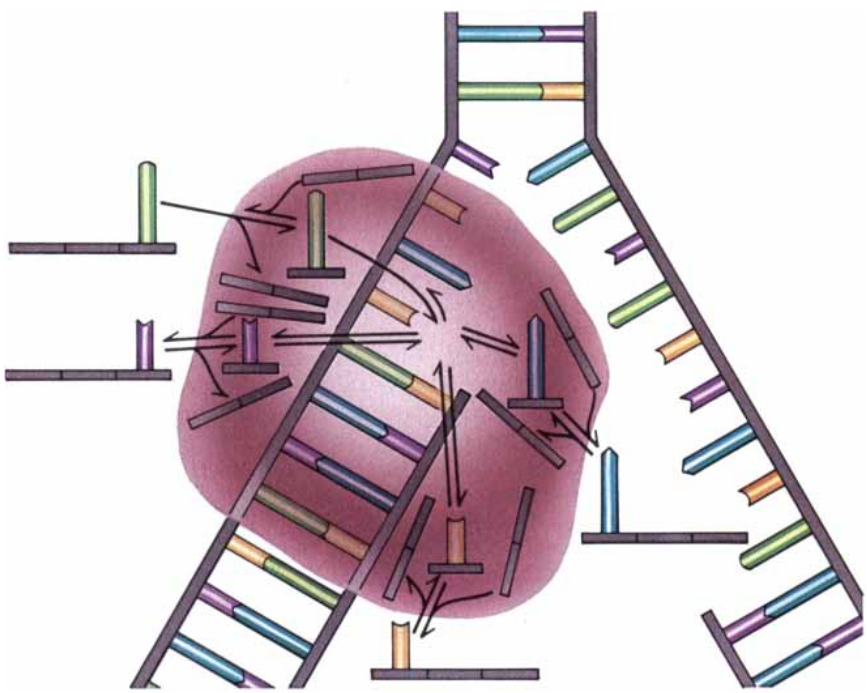
Under ordinary circumstances the combination of nucleotide selection and proofreading by exonuclease results in an error rate of about one mistake per 10 million base pairs. But both error-avoidance mechanisms can be impaired if the pool of triphosphates that supplies the raw material for synthesis has unequal proportions of the four kinds of nucleotides. In the early 1980's Alan R. Fersht and his colleagues at the Imperial College of Science and Technology in London found that the error rate at a given position on the nascent DNA strand is directly proportional to the concentration of noncomplementary nucleotides as well as to the concentration of the nucleotide that is complementary to the template base next in line. It is not surprising that a high concentration of noncomplementary nucleotides increases the likelihood of creating a mismatch, and a high concentration of the nucleotide complementary to the next base in line would hasten the polymerizing reaction, shortening the time the exonuclease has to act. Hence intracellular nucleotide pools must be delicately balanced.

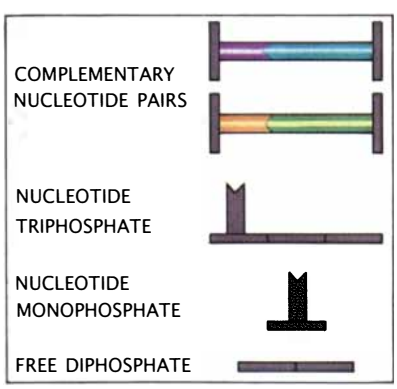
Error-avoidance mechanisms are straightforward enzymatic reactions in which more energetically desirable outcomes prevail over less stable outcomes. Error correction is a little more complicated. In order to correct a mismatch in newly synthesized DNA, the enzymatic machinery must be able to detect and remove a mismatched nucleotide, and to regenerate the correct sequence.

The phenomenon of mismatch repair was proposed almost 25 years ago by Robin Holliday of the Medical Research Council in London. For quite a long time molecular biologists puzzled over how a repair system could distinguish the parental DNA strand from the newly synthesized strand. The distinction seemed necessary if the nonparental member of the mismatch was to be excised and the parental nucleotide preserved. In 1975, during a conference in Scotland, one of us (Radman) was discussing the problem of strand discrimination with Matthew S. Meselson of Harvard University. Meselson speculated, "If I were the mismatch-repair enzyme, I would look for DNA methylation following replication."

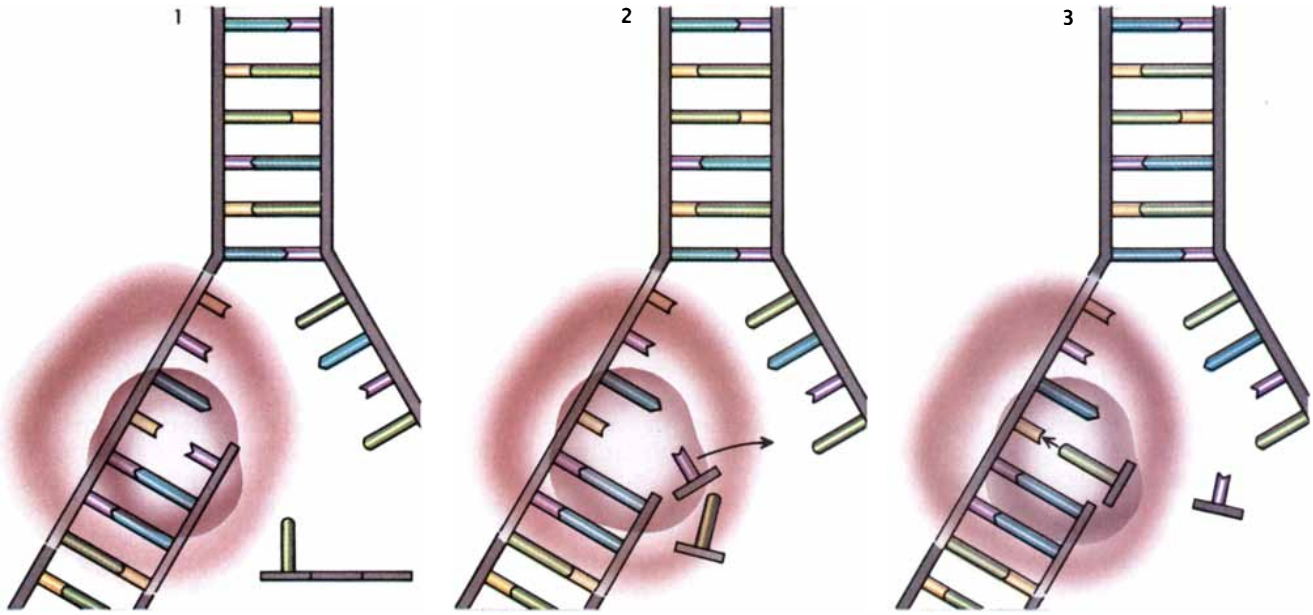
Meselson was referring to the process by which a newly synthesized DNA strand is tagged with methyl groups. In some bacteria a methyl group is attached to adenine in the sequence *GATC* wherever that sequence occurs. There is a small time lag after synthesis during which the new *GATC* sequences remain unmethylated. During that period parental and newly synthesized strands differ with respect to the extent of methylation. Meselson thought the mismatch-repair enzymes might operate within that time lag, exploiting the transient difference between young and old strands.

In 1976 the two of us met in Meselson's laboratory to test the methylation hypothesis. The results of our experiments suggested that mismatch repair does indeed occur preferentially on unmethylated DNA strands. Between 1980 and 1985 Christiane Dohet in our laboratory in Paris and Meselson's co-workers at Harvard did a series of experiments that confirmed and extended the results of our research. These workers constructed heteroduplexes containing mismatches and introduced them into various *Escherichia coli* mutants. (A heteroduplex is a DNA molecule in which the two strands do not carry identical information, because of nu-





NUCLEOTIDE SELECTION avoids the incorporation of noncomplementary nucleotides in the growing DNA strand. It is mediated by DNA polymerase (*brown*) that carries out synthesis. Synthesis involves two steps: first the triphosphate nucleotide is cleaved to a monophosphate, and then the monophosphate is added to the end of the new strand. The authors think selection acts on both steps through reversible chemical pathways (*double arrows*). If the polymerase gets hold of a noncomplementary nucleotide, the nucleotide can be rejected before it forms a bond with the strand.



"PROOFREADING" is done by an exonuclease activity that is associated with the polymerase. Exonuclease removes nucleotides that have just been added to a strand. When a mismatch at the terminal base pair slows the incorporation of the subsequent nucleotide (*1*), the exonuclease has time to act (*2*). The polymerase then repeats its search for the proper match (*3*).

cleotide mismatches, deletions, additions or damage.) They found that when one strand of a mismatch-containing heteroduplex is fully methylated, repair occurs exclusively on the unmethylated strand. If all the *GATC* sequences on both strands are methylated, no repair takes place.

Later our colleagues Françoise Langle-Rouault and Geneviève Maenhaut-Michel at the Free University of Brussels showed that the unmethylated *GATC* sequence that directs mismatch repair can be several thousand base pairs away from the mismatch itself. Furthermore, DNA with no *GATC* sequences is just as refractory to repair as DNA with fully methylated *GATC* sequences. The same investigators also found that a break in one chain of the heteroduplex can serve the same function as an unmethylated *GATC* sequence.

Langle-Rouault and Maenhaut-Michel discovered that such heteroduplexes can be repaired by a strain of *E. coli* called *mutH,* one of four *E. coli* mutants known to be deficient in mismatch repair. The three other strains are known as *mutL, mutS* and *mutU.* Because of the mismatch-repair deficiency, some of the strains are strong mutators. It turns out that each of them lacks a protein function involved in mismatch repair. The proteins are designated by the same letters as the mutants in which they malfunction.

Langle-Rouault and Maenhaut-Michel's results suggest that the strand of DNA to be repaired is cut at an unmethylated *GATC* site, and that the MutH protein is the enzyme that cuts the strand. Their conclusions and those drawn from our earlier in vivo work have recently been confirmed at the Duke University Medical Center by Paul Modrich and his co-workers, who designed an ingenious assay for in vitro mismatch-repair activity. With this assay Modrich and his colleagues have shown that the MutH protein is in fact the cutting enzyme, and he has gone on to characterize the other Mut proteins. MutS protein is known to bind to the mismatch itself; MutL has also been implicated in mismatch recognition, and MutU is an enzyme that unwinds the two strands of a double helix, a step apparently necessary for repair to proceed. Modrich's research also revealed that another protein called single-strand binding protein is required for mismatch repair. This protein, which also takes part in DNA replication, is thought to protect and stabilize unpaired strands so that they remain accessible to the polymerase.

Although many of the specifics of the *E. coli* mismatch-repair system still have not been worked out, two models have emerged to explain its operation [*see illustration on opposite page*]. Both models propose that the MutS protein, and perhaps the MutL protein, are involved in the recognition of mismatches, and both postulate some means of long-distance communication between the mismatch and nearby unmethylated *GATC* sequences. In both of the models the MutH protein cleaves the newly synthesized strand at unmethylated *GATC* sequences, the MutU protein unwinds the strand and the section containing the mismatch is discarded. Single-strand binding protein is thought to protect the exposed template until a DNA polymerase reconstructs a complementary strand that fills the gap.

The main difference between the two models has to do with the sites of MutH cleavage. One model proposes that cleavage takes place at the two *GATC* sites flanking the mismatch; the other proposes that cleavage occurs at the mismatch and at one adjacent *GATC* sequence. The former model also has MutH cleaving after the DNA strands separate, whereas the latter predicts that MutH cleaves before the strands separate. Both theories are currently being tested, but the data amassed so far do not support one over the other.

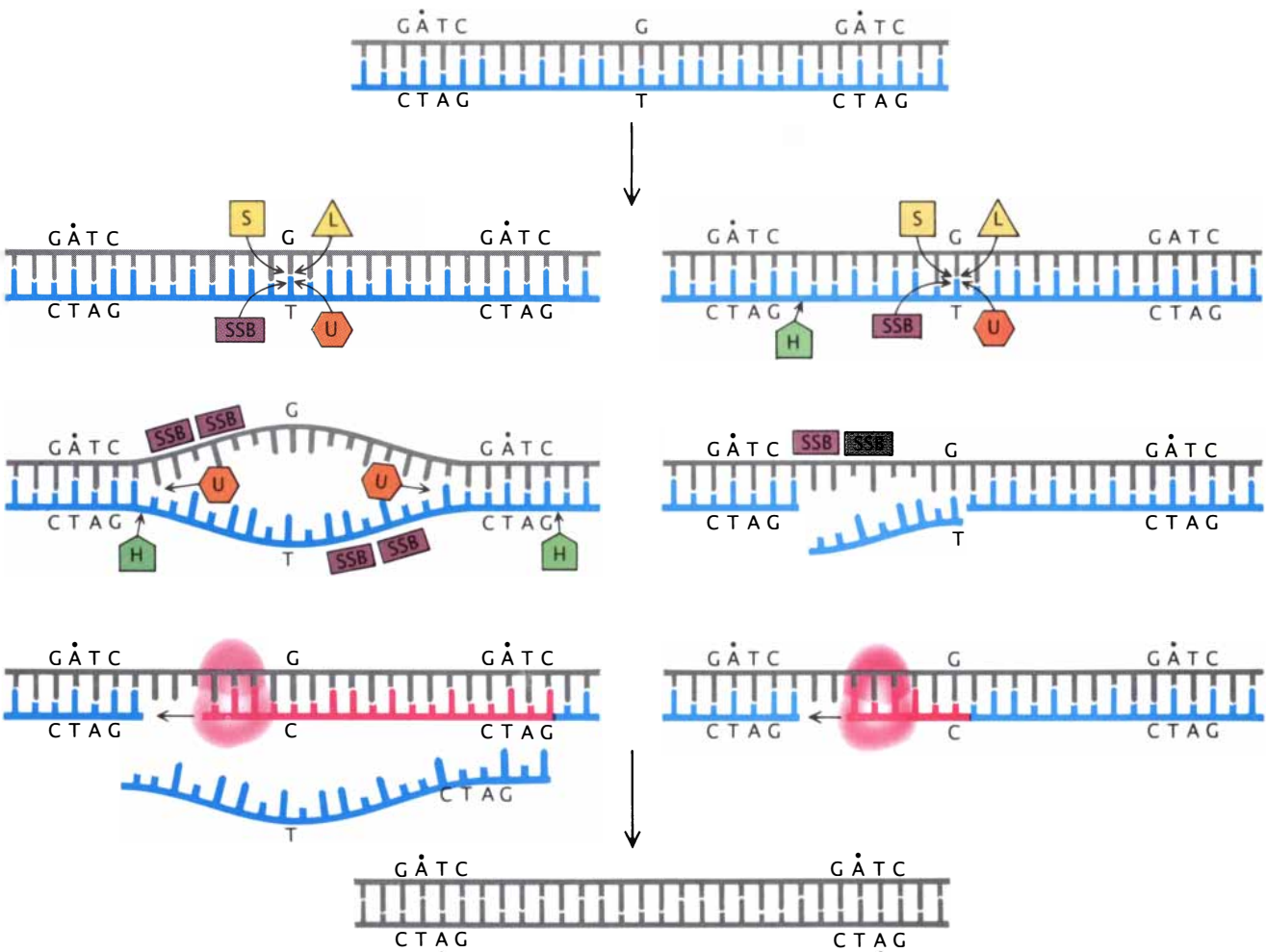
Studies completed in the past few years have also revealed something about the kinds of mismatches that are corrected by mismatch repair. In our laboratory and in the laboratory headed by Hans Joachim Fritz at the Max Planck Institute for Biochemistry in Munich the studies focused on *E.*

coli, groups led by Michel Sicard and Jean-Pierre Claverys of the University of Toulouse and by Sanford A. Lacks of the Brookhaven National Laboratory looked at mismatch repair in pneumococcus bacteria. The studies indicate that all mismatches are not repaired with equal efficiency. Instead, it seems that errors are more likely to be repaired when they do not disrupt the geometric configuration of the double helix. In other words, if a mismatch causes a bulge in the helical structure, it will probably not be detected by the mismatch-repair machinery. Such errors are beyond recall.

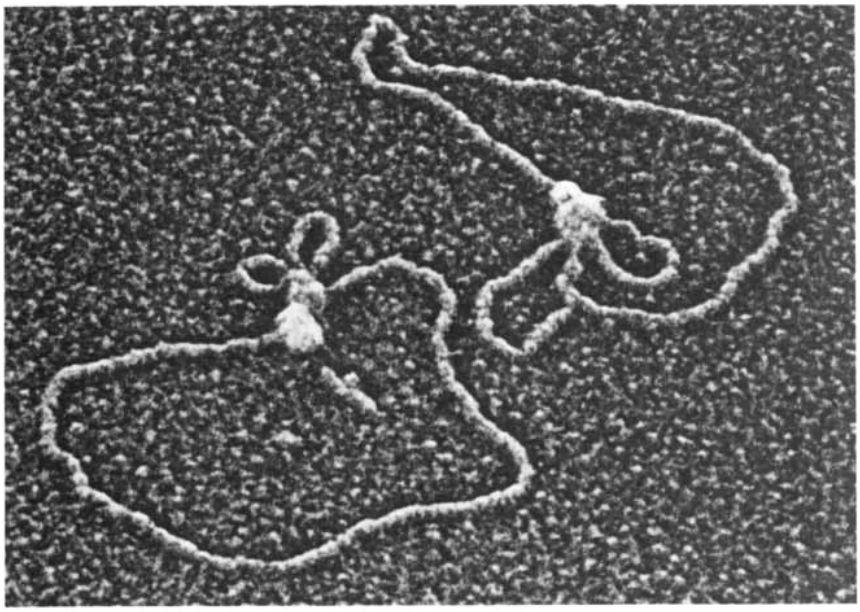
Fortunately it appears error-avoidance mechanisms are best at preventing just the kinds of mistakes mismatch repair cannot "see." Some nucleotide pairs are more likely to bulge than others, and nucleotide selection presumably creates fewest of those mismatches that cause the greatest disruption of pairing and stacking in the helix—the very mismatches the mismatch-repair system has trouble recognizing. The proofreading exonuclease also seems to be most proficient in the cases where mismatch repair is most inept.

Most of what is known about high-fidelity mechanisms of DNA replication comes from experiments with bacteria. To what extent can this knowledge be applied to more complex organisms? Both error-avoidance mechanisms are probably common to almost all organisms, where they would operate in essentially the same way. There is also ample evidence that mismatch repair occurs in yeast, fungi and fruit flies, as well as in frogs and mammals. However, most of the evidence pertains to errors arising not during replication but rather in the course of genetic recombination, in which strands of DNA are swapped between molecules of different parentage [see "Genetic Recombination," by Franklin W. Stahl; SCIENTIFIC AMERICAN, February, 1987].

One thing seems clear: since *GATC* sequences are methylated in only a handful of the bacterial species, mismatch repair cannot be directed by the methylation of *GATC* sequences in most organisms. No methylation has been detected in the chromosomes of



MISMATCH REPAIR corrects errors after a DNA strand has been synthesized. Two different models have been proposed for the mechanism of repair. In both models, proteins called MutL and MutS interact with the mismatch site (*G–T*), and a protein called MutH cleaves the newly synthesized strand. The repair apparatus distinguishes the parental strand from the new one by means of the methyl groups (*black dots*) within the parental *GATC* sequences. The strands surrounding the mismatch are separated with the help of a protein called MutU, and are stabilized by single-strand binding protein (*SSB*). The main difference between the two models has to do with where the strand containing the incorrect nucleotide is cleaved: at two flanking *GATC* sequences (*left*) or at one *GATC* sequence and the mismatch itself (*right*). In each case polymerases synthesize a new segment in place of the excised one, and the corrected strand is eventually methylated like the parental copy.



REPLICATION BEGINS on circular strands of DNA from a bacterial virus. Where the loops form knots, complexes of replication proteins from *E. coli* (enlarged 50,000 times in this image) are poised to begin copying. The micrograph was made by Jack D. Griffith of the University of North Carolina School of Medicine at Chapel Hill.

yeast and fruit flies; in mammalian cells, cytosine is methylated rather than adenine, and the methylation appears to be involved in gene regulation. How then do the cells of higher organisms distinguish the new strands from the old?

We have suggested two different answers for the question. One rests on the fact that DNA synthesis in higher organisms is initially discontinuous, so that newly synthesized strands contain transient nicks. The nicks distinguish newly synthesized strands from old strands and may provide a starting point for removing the strand containing the mismatch, much as a nick can substitute for the cutting action of the MutH protein in *E. coli.* The resulting gap can then be filled in by a polymerase.

An alternative possibility, suggested by Philip J. Hastings of the University of Alberta as well as by us, is that mismatch repair in higher organisms relies on genetic recombination. The two "sister chromatids" issuing from the replication fork have identical nucleotide sequences, and so the strands of one chromatid could become templates for reconstructing the other. In this mechanism the repair system need not distinguish between parental and newly synthesized strands; when a mismatch is encountered, both strands of the double helix could be cleaved. The cleaved strands would interact with the strands in a homologous region of the sister chromatid. The intact sister chromatid's strands would then direct the repair of the cleaved strands.

The results of experiments in our laboratory and the work of Kendric Smith at Stanford indicate that undirected mismatch repair in *E. coli* can produce double-strand breaks; Jack W. Szostak of the Harvard Medical School and Franklin W. Stahl of the University of Oregon and their colleagues have demonstrated that such breaks can be repaired by recombination. If a similar process occurs in higher organisms, it would explain the spontaneous recombination between the sister chromatids that is often observed in these organisms. One difficulty with this model is that, because mammalian cells contain multiple copies of some DNA sequences, there is a rather high risk that the cut strands will recombine improperly. Perhaps recombinational repair is restricted to the fairly short region immediately behind the replication fork by the DNA-packaging process called chromatinization.

The accuracy of DNA replication achieved by error-avoidance and error-correction mechanisms is commendable indeed. Yet they could do better. In *E. coli,* for instance, there are only a few dozen mismatch-repair proteins per cell. Why aren't there more? And why is the time during which mismatch repair must operate, as determined by the rate of methylation, kept so short?

It may be that the mismatch-repair enzymes themselves make mistakes, so that increasing the amount of such enzymes or allowing them more time to act would not improve the accuracy of replication. It could also be that mismatch repair is very costly, and that therefore a cell cannot afford to maintain more than a few dozen repair complexes.

Errors, it should also be remembered, are not always detrimental to the health of an organism. If Caesar's order had been "Execute, not liberate," a misplaced comma could have saved a life rather than cost it. Under some conditions the future of a population can depend on the ability of the individuals constituting it to mutate and thereby adapt to their surroundings. Edward C. Cox of Princeton University and Reinhard Piechocki of the University of Wittenberg in East Germany have done experiments with *E. coli* showing that when growth conditions are harsh, strains that are potent mutators dominate strains having ordinary rates of mutation. Under favorable conditions, however, excessive mutation becomes a burden and the mutator strains lose the race.

Hence the optimal efficiency of error avoidance and error correction depends on the organism and the conditions in which it finds itself. When life is good and easy, change is more a threat than a benefit; when life is hard, sometimes only change can help. As it is in so many other biological systems, the balance between flexibility and precision in DNA replication is a delicate, complex and subtle product of billions of years of evolution.

FURTHER READING

FIDELITY OF DNA SYNTHESIS. Lawrence A. Loeb and Thomas A. Kunkel in *Annual Review of Biochemistry,* Vol. 52, pages 429–457; 1982.

EFFECTS OF DNA METHYLATION ON MISMATCH REPAIR, MUTAGENESIS AND RECOMBINATION IN *ESCHERICHIA COLI.* M. Radman and R. Wagner in *Current Topics in Microbiology and Immunology,* Vol. 108, pages 23–28; 1984.

MISMATCH REPAIR IN *ESCHERICHIA COLI.* Miroslav Radman and Robert Wagner in *Annual Review of Genetics,* Vol. 20, pages 523–538; 1986.

HETERODUPLEX DEOXYRIBONUCLEIC ACID BASE MISMATCH REPAIR IN BACTERIA. Jean-Pierre Claverys and Sanford A. Lacks in *Microbiological Reviews,* Vol. 50, No. 2, pages 133–165; June, 1986.

DNA METHYLATION IN *ESCHERICHIA COLI.* M. G. Marinus in *Annual Review of Genetics,* Vol. 21, pages 113–131; 1987.

DNA MISMATCH CORRECTION. Paul Modrich in *Annual Review of Biochemistry,* Vol. 56, pages 435–466; 1987.

HELPING COMPUTER PROGRAMMERS MAKE FAST AND STEADY PROGRESS.

Until now, manual skills were a big part of writing complicated computer programs. They required a major investment in time, effort and money—and days of boredom for the applications designer.

McDonnell Douglas information systems specialists bypass the dull, turtle-speed tedium formerly necessary for good programs. Their Pro-IV® software maps out the best program path and selects routine computer instructions for them.

The result is a working program that emerges with hare-like speed, ready to use, allowing application systems designers to fully concentrate on the special characteristics of the project.

Designers like Pro-IV software because it lets them try out ideas in prototype programs, quickly and easily, selecting only the best. Managements like it because it's faster, cheaper and yields better results.

For more information, write: Software, McDonnell Douglas, Box 14526, St. Louis, MO 63178

MCDONNELL DOUGLAS

INFORMATION SYSTEMS *MILITARY & COMMERCIAL AIRCRAFT* *SPACECRAFT & MISSILES* *TRAVEL MANAGEMENT* *HELICOPTERS* *FINANCING*

Measuring Crustal Deformation in the American West

Continental crust is actively deforming as the Pacific and North America plates slide past each other. Direct measurements of the process rely on extraterrestrial reference points such as quasars

by Thomas H. Jordan and J. Bernard Minster

In several regions of the western United States, the earth's crust is actively deforming. Along California's San Andreas Fault, which is surely the most famous locus of dangerous earthquakes in North America, the deformation takes the form of horizontal slippage: crust on the western side of the fault is sliding toward the northwest relative to crust on the eastern side. East of the San Andreas Fault lies another geologically active area known as the Basin and Range Province, a 1,000-kilometer-wide region that includes the Great Basin of Nevada and western Utah. The crust in the Basin and Range Province is actively spreading apart, probably because of forces originating in the earth's mantle (the layer directly under the crust). West of the San Andreas is a much narrower geologically active zone that includes California's coastal mountain ranges and the continental margin. Here deformation manifests itself in the mountains, which are still being pushed up by complex compressive forces, and in earthquakes.

These deformations, although they may seem quite distinct from one another, can be seen as elements of a much larger motion: the relative motion of the two enormous sections of the earth's surface known as the Pacific Plate and the North America Plate. The Pacific Plate is sliding northwestward past the North America Plate at a speed of about 50 millimeters per year. The western U.S. is caught in the middle; it is there that the plate motions must be accommodated by ongoing tectonic activity: by the deformation of the earth's crust. Hence the various deformations taking place across the western U.S.—the expansion of the Basin and Range Province, the horizontal slippage along the San Andreas Fault, and the complex tectonics of coastal California—must, when seen as a whole, account for the total relative motion of the Pacific and North America plates.

How is the large-scale motion distributed across the western U.S.? How much of it is expressed in each of the three major zones of deformation?

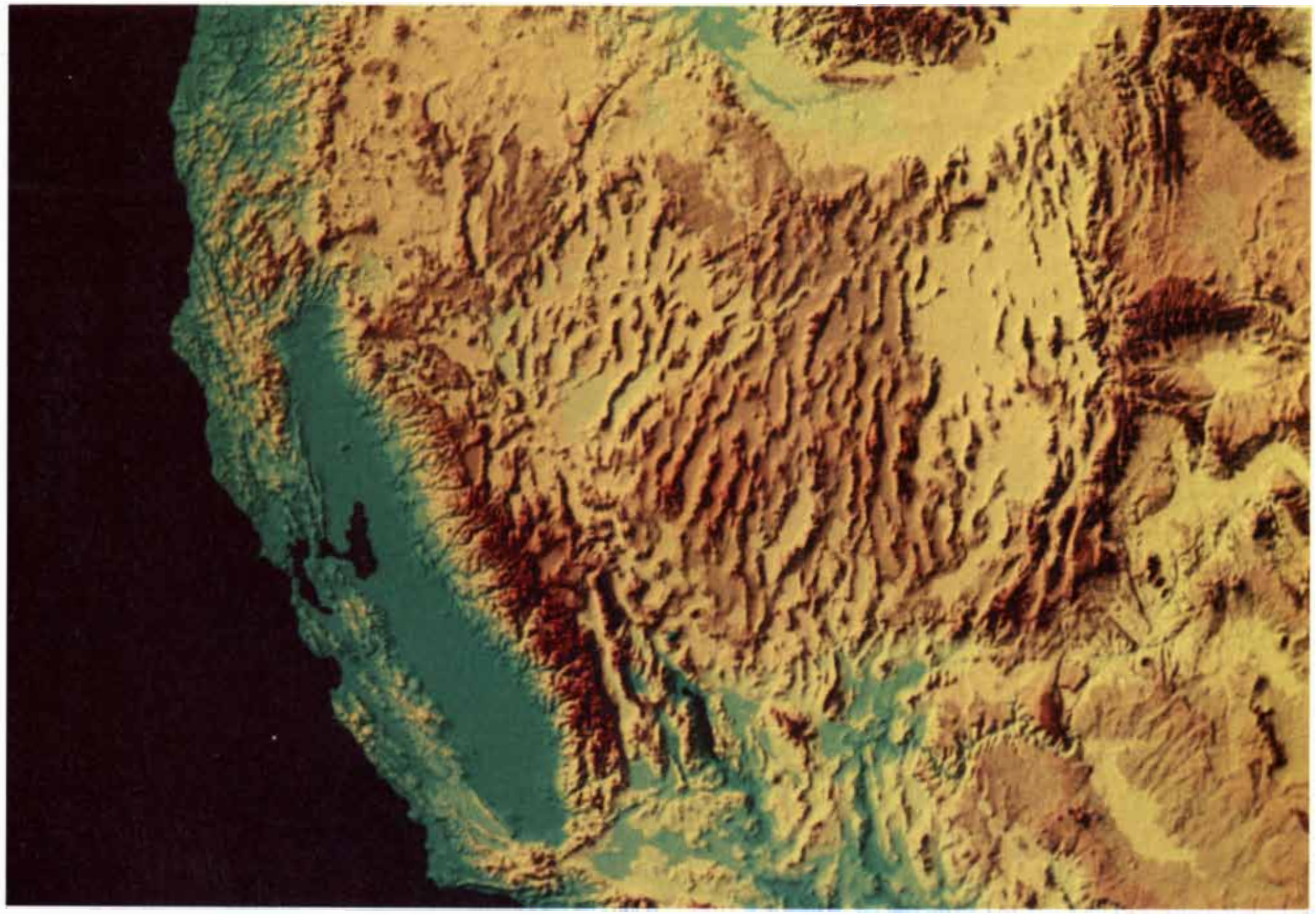
The question is interesting from a purely scientific standpoint, but it has great practical importance as well. In order to understand the risk of earthquakes in a region, geologists must first understand how the region is deforming. Unfortunately it is sometimes difficult or even impossible to observe the deformation directly. This is particularly true of coastal California, where critical geological features are submerged below the waters of the Pacific ocean; the seismic risk in the region has been the subject of intense, protracted debate. Perhaps we can improve our understanding of the deformation occurring there by viewing it as one element of the much larger picture framed by the relative motion of the Pacific and North America plates.

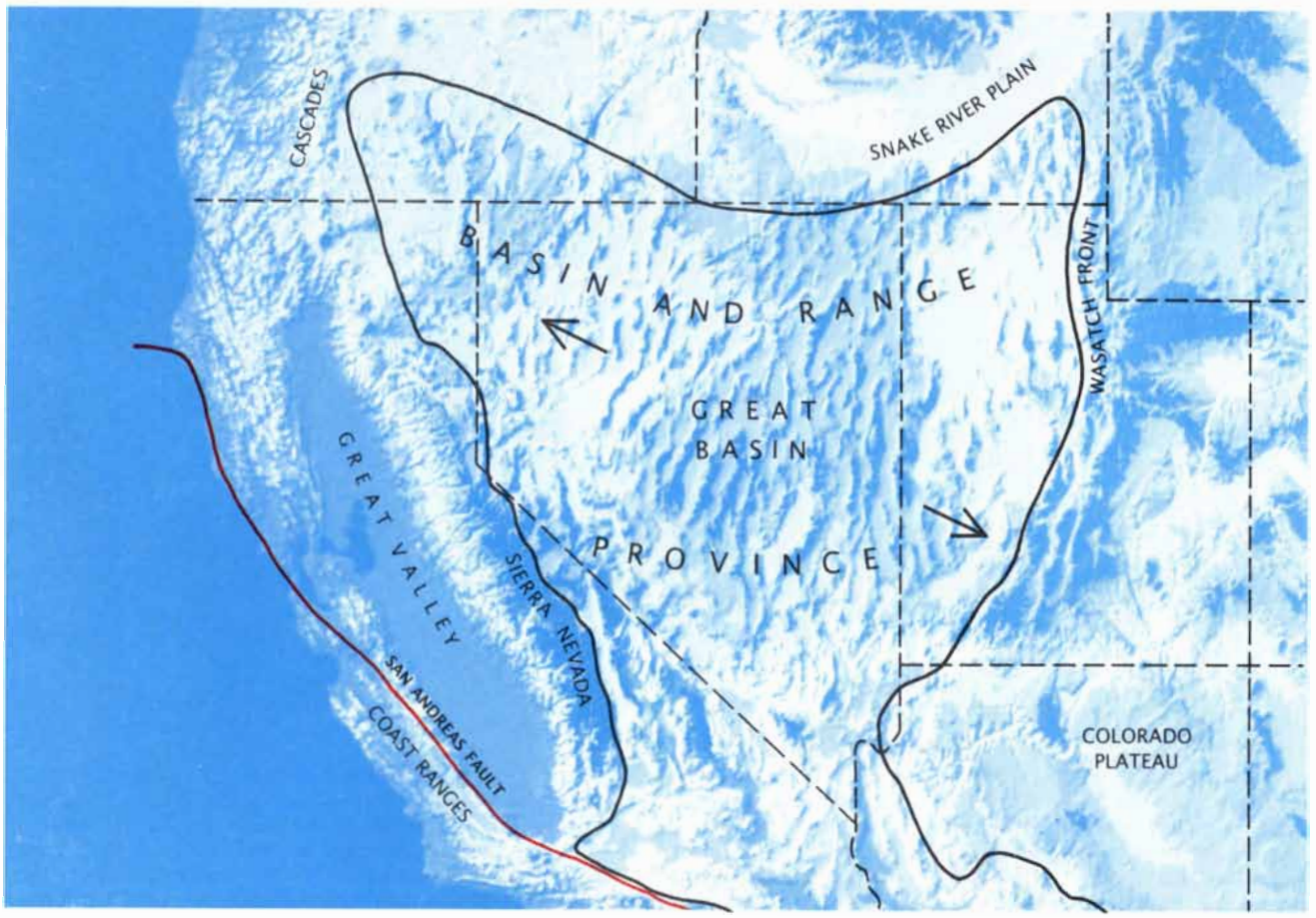
To put this picture together, we must get a better handle on the rates of deformation in the western U.S. To do so, we turn to geodesy, the science of measuring positions on the surface of the earth. Geodetic surveys employ precise observations of angles and distances to determine the relative locations of points in a network; rates of deformation can then be calculated by determining how the distances between points change over time. For example, one modern instrument, the laser Geodimeter, times the flight of a light beam bounced back from a corner reflector at a distant station to a precision of better than one ten-billionth of a second in order to measure line-of-sight distances to within a centimeter or so.

The earth's curvature and atmosphere limit such line-of-sight measurements to distances of tens of kilometers, which is sufficient for surveying narrow fault zones but not for studying the rates of deformation across regions as wide as the western U.S. Over the past two decades a new

THOMAS H. JORDAN and J. BERNARD MINSTER met as graduate students at the California Institute of Technology in 1969, and they have been working together on problems relating to seismology and tectonic deformation ever since. Jordan is professor of geophysics at the Massachusetts Institute of Technology and head of the department of Earth, Atmospheric and Planetary Sciences there. He went to M.I.T. in 1984 after working at Princeton University and at the Scripps Institution of Oceanography. Minster is a visiting professor at Scripps. From 1976 to 1982 he taught at Caltech, and he has also worked at S-Cubed in La Jolla, Calif., and at Science Horizons, Inc., in Encinitas, Calif.

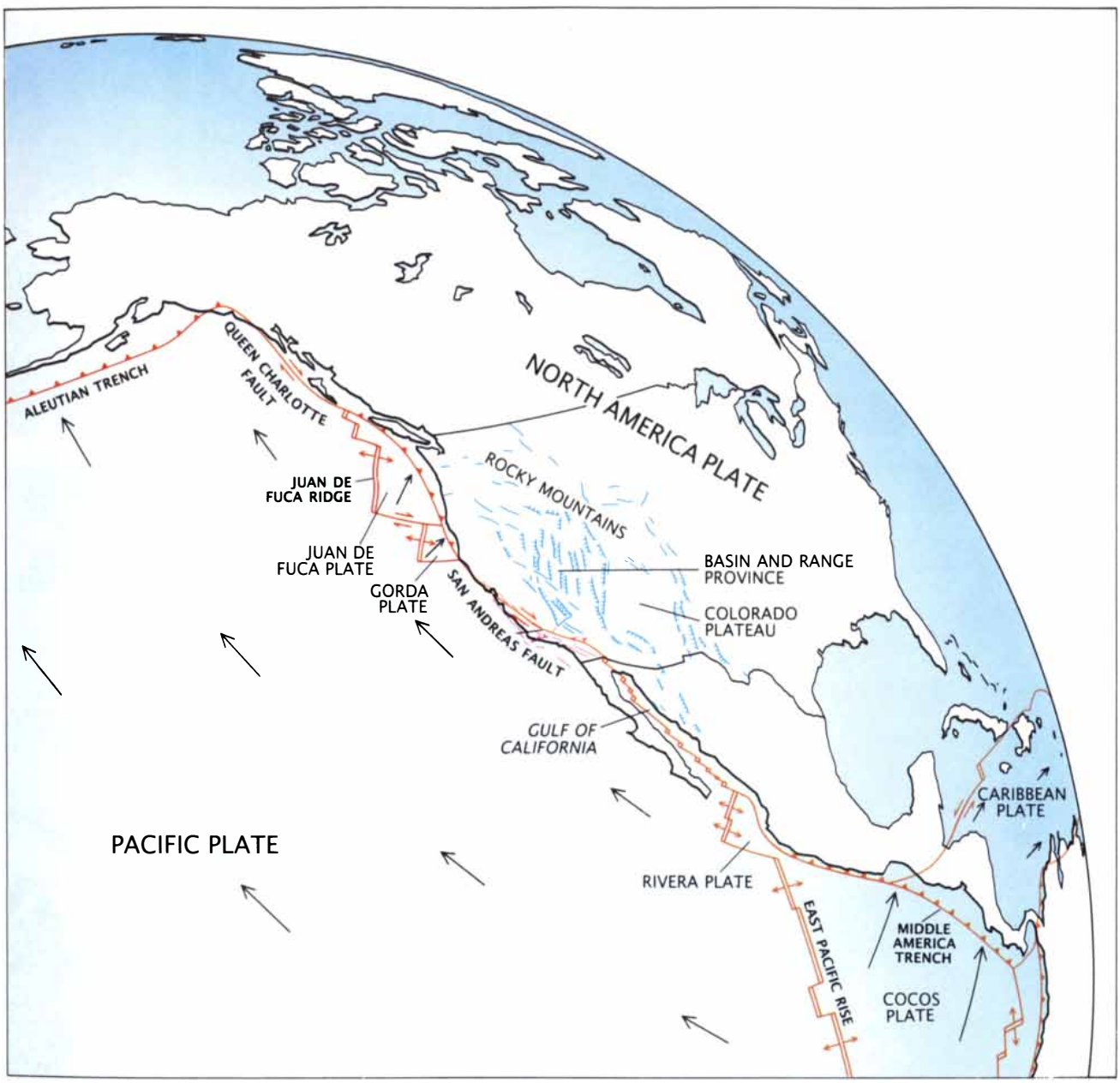
WESTERN U.S. is deforming actively in several regions. The Great Basin (*center*) is expanding. Features caused by the expansion include ridges called horsts and valleys called grabens, which run roughly perpendicular to the direction of expansion. Along the San Andreas Fault (*left*) the deformation takes the form of horizontal slippage: crust west of the fault is sliding to the northwest relative to crust east of the fault. Coastal California (*bottom left*) is also deforming; it is being reshaped by earthquakes and by folding of the crust, creating new mountain ranges. The upper image was generated by computer from a data base of more than eight million topographic points compiled by the National Oceanic and Atmospheric Administration.





set of techniques have been developed that make use of extraterrestrial reference points—earth-orbiting satellites and extragalactic radio sources such as quasars—to monitor deformation. These "space-geodetic" techniques make it possible to measure the distance between points 1,000 kilometers apart to a precision of a centimeter or less. The technology of space geodesy is new, but it has revolutionized the science of distance measurements and is already contributing direct observations of geological motions in the American West.

The suggestion that there is a close relation between deformation in the western U.S. and the large-scale motions of tectonic plates has its origins in studies of the San Andreas Fault. In the famous 1906 San Francisco earthquake a 420-kilometer-long section of the San Andreas ruptured, causing extensive damage and triggering a disastrous fire in the city. Fences, roads and railway tracks running across the fault were cut and offset, so that segments on one side of the fault did not meet those on the other side. Careful observations of this phenomenon led Harry Fielding Reid of Johns Hopkins University to formulate what is now known as the elastic-rebound theory of earthquake faulting [see "The Motion of the Ground in Earthquakes," by David M. Boore; SCIENTIFIC AMERICAN, December, 1977]. According to this theory,



RELATIVE MOTION of the enormous sections of the earth's surface called the Pacific Plate (*left*) and the North America Plate (*right*) is the primary driving mechanism for earthquakes and deformation in the western U.S. Along the west coast of North America the Pacific Plate moves northwestward at a rate of about 50 millimeters per year relative to the North America Plate. About two-thirds of the motion is reflected in horizontal slippage along the San Andreas Fault. The remaining motion is expressed as extension in the Basin and Range Province of Utah, Nevada and Arizona, as horizontal slippage along faults in California other than the San Andreas, and as other deformation (such as mountain building) in coastal California.

the fault separates blocks of crust that are moving at a steady rate with respect to each other; friction within the fault prevents the blocks from slipping until the accumulated stress exceeds the strength of the rocks, when an almost instantaneous displacement—an earthquake—occurs.

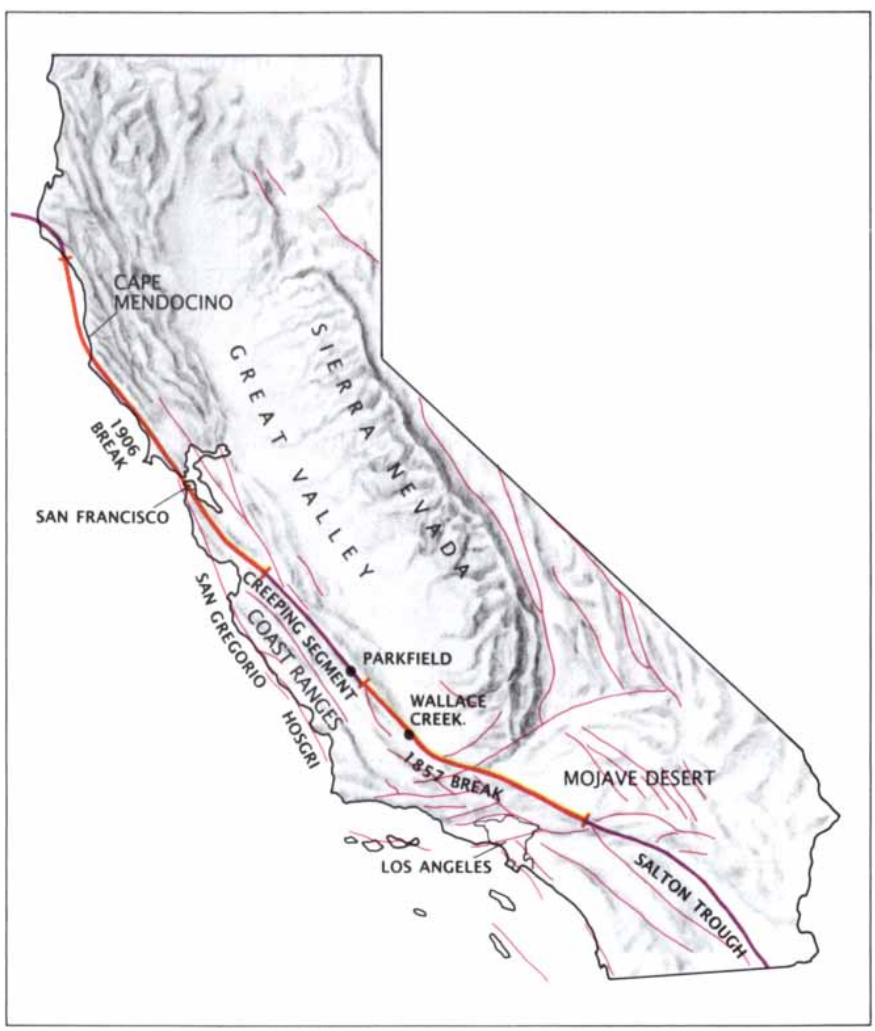
In a landmark paper written in 1965, J. Tuzo Wilson, then at the University of Toronto, suggested that the slippage along the San Andreas is actually driven by the relative motion of two of the large plates into which, as geologists were then realizing, the earth's surface is divided. The geophysicists of the late 1960's, who were trying to develop the first self-consistent models of global plate motions, adopted Wilson's suggestion that the San Andreas is the boundary between the Pacific and North America plates.

At that time direct estimates of the present-day rate of slip along the fault were not available, but analysis of the rate at which new crust has been created in the Pacific ocean (an indicator of plate speeds) suggested that the plates were moving relative to each other at a rate of about 60 millimeters per year. Geologists were quick to point out, however, that if the rate of slip along the San Andreas Fault had been constant at this high value, certain volcanic flows that were continuous across the fault at the time of their formation—about 10 million years ago—should now be offset by 600 kilometers; the flows are actually displaced by only 250 kilometers.

On the other hand, motion along the San Andreas has not proceeded at a constant speed for 10 million years, and there have been drastic changes in the plate geometry over that time. Attempts were made to reconcile the low geological rates with the high plate-tectonic rates by appealing to these temporal changes. Perhaps, for example, the present-day rate of slip on the San Andreas is higher than the geological rate. In 1973, however, James C. Savage and Robert O. Burford of the U.S. Geological Survey showed that there is indeed a discrepancy between the rate of horizontal slippage on the San Andreas and the relative motion of the Pacific and North America plates: precise Geodimeter surveys in central California revealed that between 1960 and 1970 the rate of slip on the San Andreas was only 32 millimeters per year, with an uncertainty of only five millimeters per year.

At about the same time, uncertainties in the description of plate motions were being reduced by the derivation of plate-tectonic models that made systematic use of the vast array of data collected by oceanographic institutions over the previous decade. We, in collaboration with Peter H. Molnar, who was then at the Scripps Institution of Oceanography, constructed a model of plate velocities averaged over the past two to three million years called RM1 (for relative motion 1). RM1 indicated that the relative speed of the Pacific and North America plates was roughly 57 millimeters per year, which was consistent with the earlier estimates and almost twice the rate of horizontal slippage found by geodesy across the San Andreas Fault. Molnar and Tanya M. Atwater, also at Scripps, noted that, in the absence of geological evidence for major variations in the slip rate over the past few million years, the most obvious way to explain the difference between the geodetic and plate-tectonic values was to ascribe the "missing motion" to displacement on faults other than the San Andreas, specifically those known to be active in the Basin and Range.

In the 15 years since the "San Andreas discrepancy" was first recognized, the observations that made it apparent have been augmented and improved. Savage and Burford's original estimate of the rate of horizontal slip along the San Andreas has held up well. In addition geological evidence—detailed mapping and dating of soils laid down by a creek that runs across the fault, carried out by Kerry E. Sieh of the California Institute of Technology and the late Richard H. Jahns of Stanford University—showed that the average rate of slip for the past 13,000 years has been 36 millimeters per year, about the same as the current rate. The close agreement with the geodetic findings is significant, because the geological data reflect the cumulative effects of many large earthquakes and are therefore fair-



ACTIVE FAULTS, along which earthquakes are likely to occur, are abundant in California. The most famous is the San Andreas Fault (*dark red and purple*), on which many major earthquakes have taken place. Other active faults are shown in red.

ly immune to short-term fluctuations.

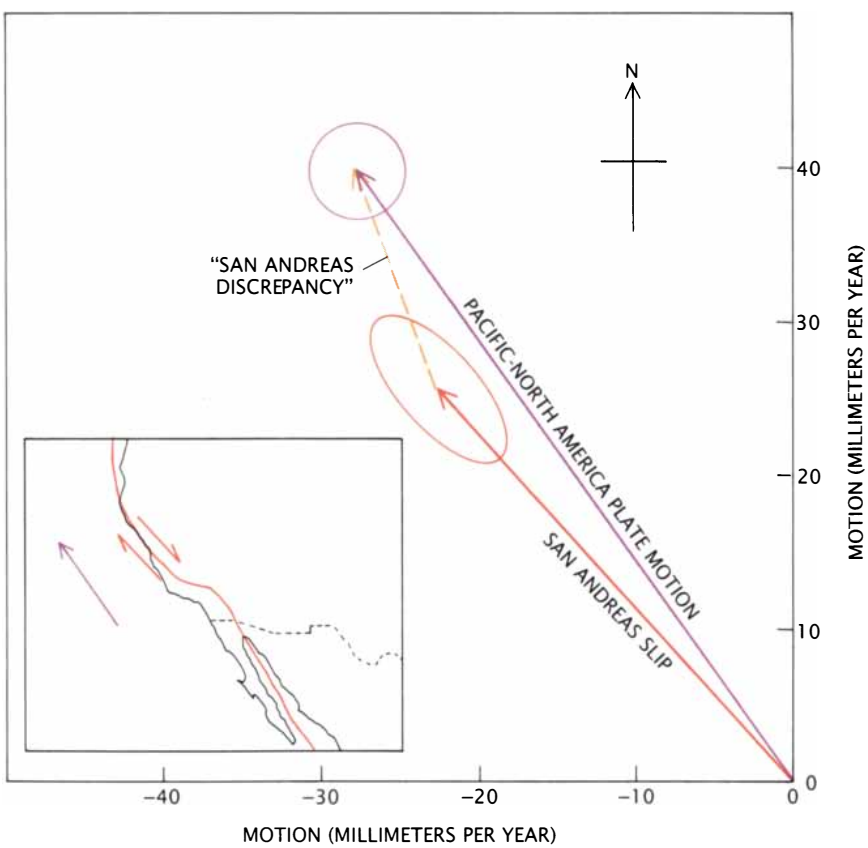
Estimates of the tectonic-plate velocities have been improved as well. In 1978 we compiled an improved set of oceanographic data to produce an updated model of global plate motions called RM2. Although RM2 included substantial revisions to the plate velocities calculated for many parts of the world by RM1, the value it implied for the rate of relative motion of the Pacific and North America plates—56 millimeters per year with an uncertainty of three millimeters per year—was essentially the same as the RM1 value. Nearly the same speed was estimated in another global model published in the same year by Clement G. Chase of the University of Minnesota.

The discrepancy between the rate of slip observed at the San Andreas Fault and the calculated rate of plate motions thus remained about 22 millimeters per year—a whopping 40 percent of the plate motion! How could this discrepancy be explained? Since 1978 investigators have focused on three primary possibilities: that the RM2 estimate of plate motions is too high, that the missing motion is accounted for by the spreading of the Basin and Range Province and that the missing motion is accounted for by deformation in coastal California. These hypotheses are not mutually exclusive; as it turns out, a combination of all three fits the data best.

The RM2 estimate was inherently uncertain because of the difficulties involved in estimating the motions of the Pacific and North America plates. An understanding of the relative velocity of the plates depends critically on the interpretation of the sea-floor geology of a tiny segment of an underwater ridge in the mouth of the Gulf of California just north of a fault called the Tamayo Transform.

When RM2 was devised, very few data were available from this region. The area has been surveyed several times since, and from the new data Charles DeMets and his colleagues at Northwestern University found in 1987 that the plates are moving relative to each other at a rate of about 48 millimeters per year. This new figure reduces the San Andreas discrepancy to roughly 15 millimeters per year, a smaller, although still considerable, value.

Where is this missing motion? The shortfall between the rate of slip on the San Andreas and the total motion of the plates is enough to make a seismologist or a public-safety official nervous. It is comparable, for example, to the total slip rate across the North Anatolia Fault of Asia Minor or the Motagua Fault of Guatemala, both of which have long histories of killer earthquakes. The missing motion could be spread out in many faults across the western U.S., it could be concentrated in a few faults or it could be taken up by some steady, aseismic deformation of the crust. These possibilities clearly have very different implications for the risk of earthquakes. One way to begin attacking the problem is to ask how much of the missing motion takes place in regions to the east of the San Andreas Fault, such as in the Great Basin, and how much takes place to the west of it, along the California coast.

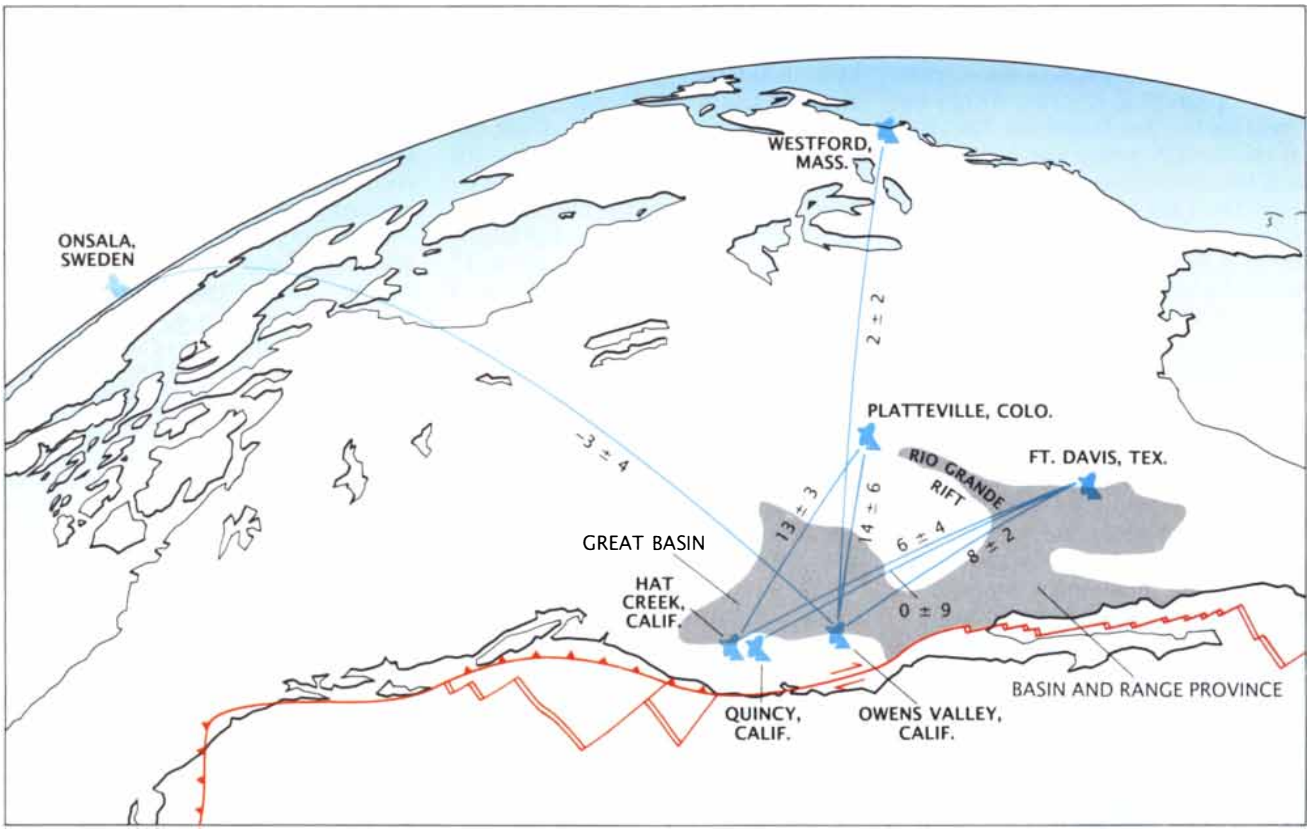


"SAN ANDREAS DISCREPANCY" is the mismatch between the rate and direction of horizontal slippage along the San Andreas Fault and the relative motion of the Pacific and North America plates. Purple vector indicates how the Pacific Plate moves with respect to the North America Plate. Dark red vector indicates the speed and direction of slippage along the San Andreas. Broken orange vector is the motion that must be taken up somewhere in the western U.S. other than at the San Andreas.

Our tools for analyzing the missing motion are vectors. The vectors, drawn as arrows, represent motions; the length of a vector represents the rate of motion, and the direction in which the vector points represents the direction of motion. The net effect of any combination of motions can be determined by a geometric process known as vector addition: vectors representing the motions are placed head to tail, and another vector is drawn joining the tail of this assembly to its head. This vector, called the vector sum of the assembly, represents the net motion.

In our case we know that the net motion—the sum of the contributions due to slippage on the San Andreas, to expansion east of the San Andreas, in the Basin and Range, and to deformation west of it, in coastal California—should be equal to the relative motion of the Pacific and North America plates. We therefore know that vectors representing these three kinds of motion, when placed head to tail, should add up to a vector equal in length and orientation to a vector representing the relative motion of the plates.

As a reference point for our vector calculations we have chosen a point located on the San Andreas Fault at a latitude of 36 degrees north, a few miles north of the small town of Parkfield. (Parkfield is small indeed—the hamlet has a population of fewer than 50—but it is famous in seismological circles because it has been predicted that an earthquake will occur there



SPREADING OF THE GREAT BASIN is gauged from changes in the length of baselines measured by very-long-baseline interferometry (VLBI). In VLBI radio telescopes record high-frequency "noise" from a number of extragalactic sources (such as quasars). The distances between telescopes can be computed to within a centimeter or so by comparing the arrival times of the radio waves. Changes in the baselines, measured between 1980 and 1985 by the National Aeronautics and Space Administration's Crustal Dynamics Project, are given in millimeters per year. The rate of change along the baseline between Onsala and Owens Valley has been corrected for the relative motion of the North America and Eurasia plates.

within the next five years.) In this part of California the fault trends in a direction 41 degrees west of north. A patient geodesist stationed just east of the fault zone near this reference point will observe that a marker placed to the west of the fault drifts to the right (that is, to the northwest) with a time-averaged velocity of about 34 millimeters per year.

This information—the direction and rate of slip—enables us to draw the vector representing motion due to slippage along the fault. The next step is to find the direction and speed of the motion occurring east of the San Andreas, which is primarily associated with the extension of the Basin and Range. Then it is a relatively simple matter to find the remaining vector, representing the motion due to deformation in coastal California. We simply add the vectors representing horizontal slippage along the San Andreas and extension in the Basin and Range, and compare the sum with the vector representing the relative motion of the Pacific and North America plates, which we derive from plate models such as RM2. Any difference is probably accounted for by deformation in coastal California. Then we shall have a quantitative description of how the motion of the plates is distributed among the three major domains of deformation in the western U.S.

The geological investigation of the Basin and Range began with the work of John Wesley Powell, Grove Karl Gilbert and other pioneers of the U.S. Geological Survey, who recognized the basic pattern of the deformation nearly a century before the discovery of plate tectonics. The details are complex, but the dominant element of the deformation is extension along a line running roughly northwest-southeast. As the crust spreads, some blocks of crust sink, forming grabens: valleys that trend roughly perpendicular to the direction of spreading. Ridges left between the grabens are called horsts. At the midlatitudes of Denver and San Francisco, the Basin and Range includes about 20 horst-and-graben structures, extending from the Wasatch Front (a major fault system running roughly north-south through Salt Lake City) to the two-mile-high scarp of the Sierra Nevada. Most of this expanse is occupied by the Great Basin, a closed, 500-kilometer-wide depression formed by the stretching and thinning of the crust.

In this region there are a number of ways to find the net direction of expansion. The horst-and-graben structures lie nearly perpendicular to the direction in which blocks of crust move as they spread apart, and so the record of slippage found on the exposed scarps gives a good indication of the direction in which the region has stretched. Additional data come from analysis of certain volcanic formations. For example, cinder cones tend to form in lines along vertical cracks that are perpendicular to the direction of extension. The direction in which the crust is being stretched can also be inferred by drilling deep boreholes and pumping in fluids under high pressure; the crust tends to fracture on planes perpendicular to the axis of greatest tension.

In 1980 Mark L. and Mary Lou Zoback

of the U.S.G.S. combined a number of these techniques to produce a comprehensive description of the orientations of stress and strain within the Great Basin. They found that the crust to the west, between the Great Basin and the San Andreas Fault, is moving away from the relatively stable North American continent east of the Great Basin in a direction about 60 degrees west of north. This, then, is the direction of the vector representing motion due to extension in the Great Basin.

Not surprisingly, estimating the rate of extension (which will determine the length of the motion vector) from geological data has proved to be more difficult than estimating its direction. In a review published several years ago we found geological and geophysical support for rates ranging from as little as one millimeter per year to more than 20 millimeters per year.

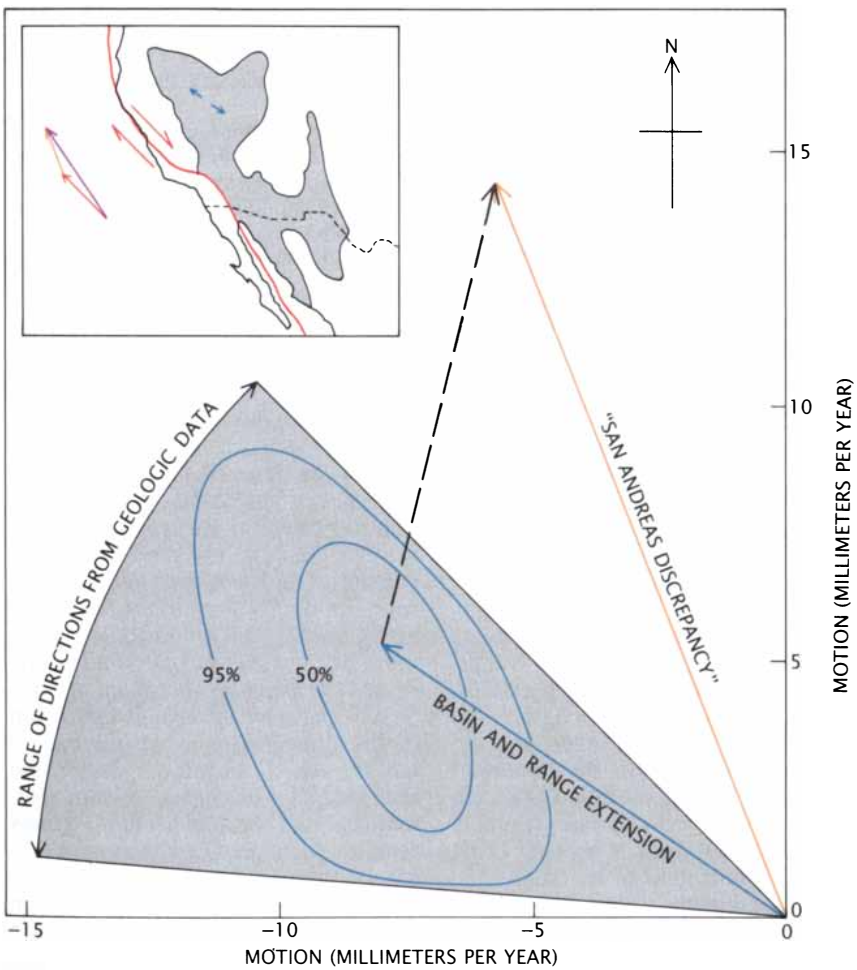
The best geological rate estimates have come from the new mapping techniques of paleoseismology, a budding branch of geological study that focuses on the displacements caused by individual prehistoric earthquakes. To date few faults in the Basin and Range have been studied in detail, but based on preliminary mapping by Robert E. Wallace of the U.S.G.S. we have concluded that the rate of extension across the Great Basin has probably not exceeded an average of 12 millimeters per year over the past 12,000 years. Unfortunately some of the assumptions that go into estimating the rate of extension from geological and seismological observations are questionable, the data redundancy is low and our confidence in these rate estimates is not very high.

Geodesy, on the other hand, should be able to provide a direct, independent and accurate measurement of the extension rate. Although line-of-sight geodetic measurements are limited to distances less than a few tens of kilometers, the northwest-southeast extension of the Basin and Range Province could, in principle, be determined by repeated measurement of a network spanning the entire province. This would be a tedious and expensive operation, however, requiring the survey of many intermediate reference marks. Moreover, because errors would accumulate at every step, the end result would not be accurate enough for tectonic modeling.

The alternative is space geodesy. One particularly fruitful space-geodetic technique, known as very-long-baseline interferometry (VLBI), is in principle quite simple [see "Studying the Earth by Very-Long-Baseline Interferometry," by William E. Carter and Douglas S. Robertson; SCIENTIFIC AMERICAN, November, 1986]. The high-frequency "noise" emanating from a quasar is recorded by radio telescopes at separate stations and cross-correlated to determine the difference between each radio wave's arrival times at the stations. These time differences measure directly how much closer one station is to the source than the other. By making a series of measurements for a set of quasars spread across the sky, one can find the position of each station in the reference frame of the quasars and determine how that position changes with time.

Although VLBI was pioneered in the 1960's, the technique did not achieve the one-centimeter precision required for studies of crustal deformation until about 1980. By that time a global network of VLBI stations had been established by the National Aeronautics and Space Administration under the auspices of its Crustal Dynamics Project, and observations of interplate motions were soon forthcoming. Within five years teams led by Thomas A. Clark of NASA's Goddard Space Flight Center and by Thomas A. Herring and Irwin I. Shapiro of the Harvard-Smithsonian Center for Astrophysics were able to collect and process enough VLBI data to estimate rates of motion along baselines spanning both the Atlantic and the Pacific oceans; the rates generally differed from the calculations of plate-tectonic models such as RM2 by only a few millimeters per year.

Spurred by this success, we have used VLBI data to estimate both the direction and the rate of extension across the Basin and Range Province. The estimated rate of extension is nine millimeters per year, with an uncertainty of four millimeters per year, and the direction is 48 degrees west of north, plus or minus 17 degrees. These geodetic values are indepen-



VECTOR REPRESENTATION of extension in the Basin and Range (*blue*) shows the motion is too slow and in the wrong direction to account completely for the San Andreas discrepancy (*orange*). The rate of extension was found largely from VLBI data, and the direction primarily from geological data. Ellipses define the range of vector lengths and directions within the indicated levels of observational confidence.

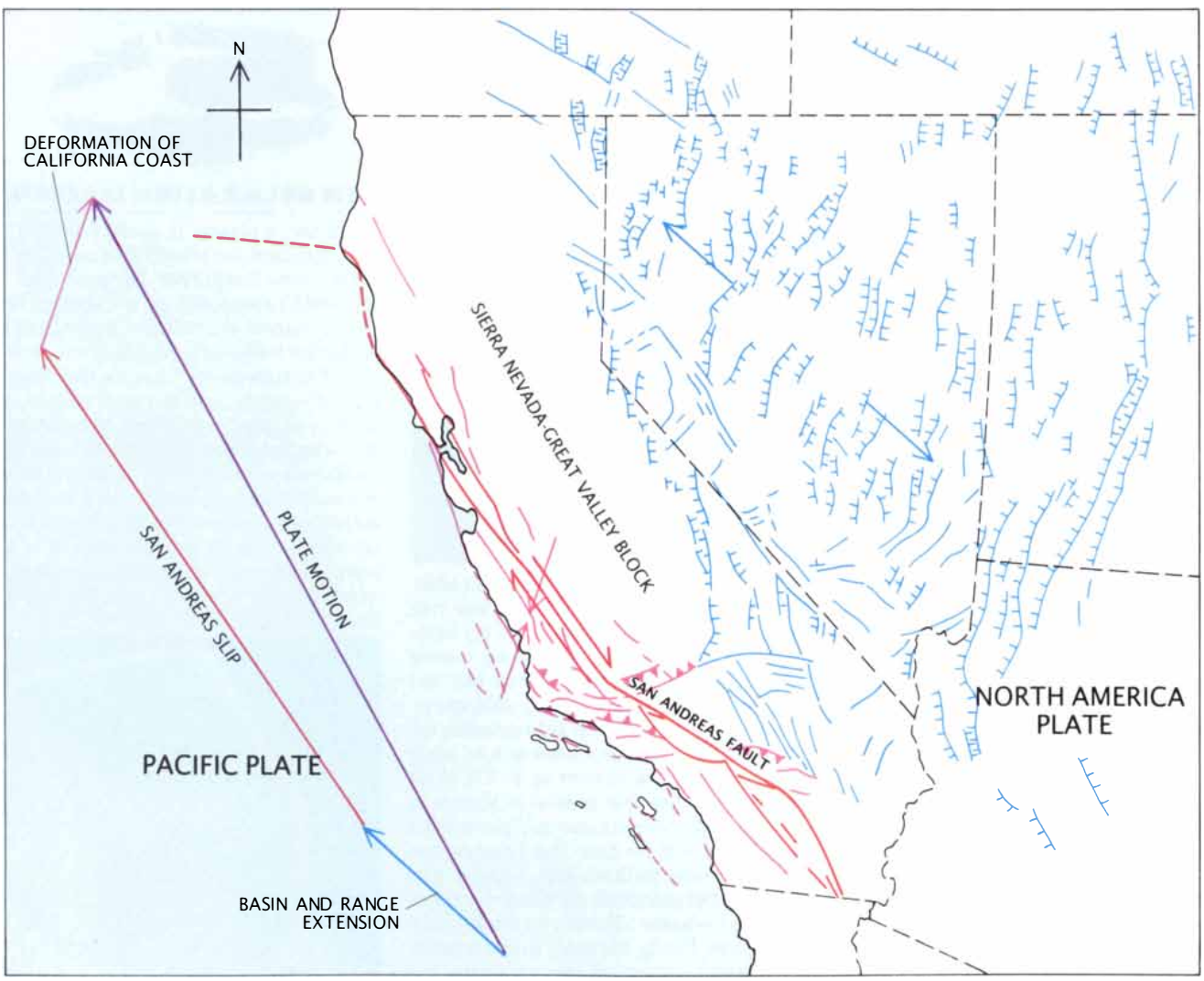
dent of the geological data and consistent with them. The VLBI data provide a more precise rate, and the geological data provide a better direction; in calculating the vector representing extension in the Great Basin, we have therefore relied mainly on VLBI data for the vector's length and on geological data for its direction.

The vector we obtain by combining the data sets indicates an extension rate of 10 millimeters per year, plus or minus two millimeters per year, and a direction 56 degrees west of north, plus or minus 10 degrees. This is too slow and in the wrong direction to account completely for the San Andreas discrepancy. By the simple process of vector addition, we find that deformation in coastal California results in a net motion of about nine millimeters per year in a direction about 14 degrees east of north.

This, then, is how the plate motion is distributed across the western U.S.: about 15 percent of it is accommodated by extension in the Basin and Range, about 70 percent is accommodated by slippage along the San Andreas Fault, and the remaining 15 percent is taken up by other deformation in coastal California. As our vector diagram shows, not all of the motion in these three zones of deformation is parallel to the plate motion. Both the extension of the Basin and Range and the deformation of coastal California include significant components of motion perpendicular to the plate motion. These components balance each other almost perfectly: they are almost equal in magnitude and point in opposite directions. This geometric balance suggests it is the expansion of the Basin and Range Province that is driving compression in the California coast ranges.

How does our vector diagram help us to understand the nature of the deformation taking place in coastal California? After all, a vector describes only the overall motion associated with deformation; it provides very little information about the details of the deformation itself. One way to analyze the estimated motion is to resolve it into its two components: a shear component, parallel to the general trend of the San Andreas Fault, and a compressional component, perpendicular to the San Andreas. We estimate that the rate of shear motion is about six millimeters per year and the rate of compression is about eight millimeters per year.

The presence of both components is consistent with what is known about the tectonics of California, which has many active faults other than the San Andreas. The existence of a compressional component is indicated by the folding and thrusting taking place in



VECTOR ADDITION indicates how relative motion of the Pacific and North America plates is partitioned among domains of deformation in the western U.S. The vector diagram shows the relative rate and direction of plate motion (*purple*), extension in the Basin and Range (*blue*), slippage along the San Andreas (*dark red*), and deformation in coastal California (*red*).

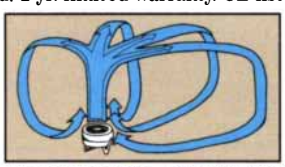
THE LIFESTYLE RESOURCE™

At The Lifestyle Resource we give you all the facts and details necessary to make an informed purchase. Your satisfaction is our primary concern. If your purchase doesn't meet your expectations, return it in original condition within 30 days for prompt refund.

AERODYNAMIC COOLING

The Turbo-Aire fan moves more *air and provides more cooling than ordinary fans* — you get more than twice the air delivered by oscillating room fans. Patented and computer-designed to maximize efficiency and minimize noise and vibration, this fan creates an exceptionally strong, smooth jetstream column that sets a room stirring with secondary air currents. Refreshes better than the intermittent blast of air from an oscillating fan. Aerodynamic housing increases blade-tip efficiency over conventional fans. Adjusts to any angle. Floor, table or wall mount. Set in the hassock position, the five bladed, 12″ fan redistributes air in an entire room, makes it as useful in winter as in summer. Can reduce air-conditioning costs. 300% more efficient than ordinary fans with the same type and size motor. 3 speed fan control, the most advanced 12″portable house fan.Weighs less than a case of cold soft drinks — and more easily moved! 2-yr. limited warranty. UL listed. **$69.95** #2160.

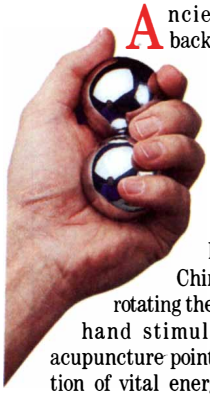
Turbo-Aire's jet stream



Hassock position — venturi effect

Conventional fan

FROM CHINA TO YOUR HEALTH



Ancient mandarins dating back 800 years believed that these Chinese Exercise Balls induced well-being of the body and serenity of spirit. These treasured gifts were given to President Reagan and his wife while visiting the Peoples Republic of China. The Chinese say that rotating the balls in the palm of each hand stimulates the fingers and acupuncture points, and improves circulation of vital energy throughout the body. Sports enthusiasts, musicians, computer users and health-conscious people everywhere consider them great muscle conditioners. Arthritis sufferers feel a decided benefit from this gentle but challenging exercise. Very effective for relaxation and meditation, Chinese Exercise Balls emit a distantly mysterious chime as you turn them. Beautifully handcrafted, 45mm. hollow polished chrome balls are perfectly weighted and fit comfortably into the average man's or woman's hand. In exquisite silk brocade box. **$29.95** #1700.

BRING MOUNTAIN TOP FRESHNESS INDOORS

The new Bionaire 700, no bigger than a table-model radio, will clean and recharge your stale indoor air to virtually mountain top freshness. Get relief from breathing allergy causing dust, pollen, tobacco smoke, animal hair and dander, cooking odors, soot, and mold spores. The Bionaire 700 will clean and rejuvenate the air in a 12x12x8 ft. room 4 times an hour, while the filtering system removes up to 99% of all particulate pollutants as small as .01 microns in the process. The filtering process begins with an activated carbon pre-filter that helps remove odors and large particles; next, with the patented electret main filter, the Bionaire removes particles as small as 1/10,000th the diameter of a human hair. Finally, Bionaire's unique negative ion generator—which not only precipitates any remaining particles, but also generates millions of negative ions—reproduces the stimulating effect of fresh mountain air. Two year limited warranty, UL listed, weighs 5.2 lbs., **$149.95** #2070. Set of two replacement filters **$19.95** #2071.

BANISH FLEAS

Flea season is the bane of existence for most dogs and cats — and their owners. Pets suffer tremendous itching and pain from these pesky little varmints, not to mention the injury animals can cause themselves from continuous scratching and biting. Now there's a safe, veterinarian-tested remedy that eliminates the use of poisonous chemicals and constant and expensive bathing, dipping, spraying and powdering your favorite pet. This clean and odorless electronic flea collar employs a pulse modulated burst circuit (PMBC) to create such a high frequency sound, inaudible to humans and pets, that the critters hastily abandon their host animal. This new and improved, water-resistant, Microtech-2® collar works on one long life lithium battery (included), and focuses on a four-foot zone of protection. Dog collar #2180 or safety breakaway collar for cats #2190; **$39.95** each.

NEW RELAXATION MACHINE

Marpac, a pioneer in sound conditioning equipment, has produced the *new* Marsona 1200A Sound Conditioner. Equipped with an improved tone generator, the new Marsona electrically synthesizes a variety of pleasing natural sounds that help mask out and reduce the annoyance of unwanted noise - it is the finest instrument of its kind. Unabated noise reduces our abilities to relax, read, sleep, concentrate or otherwise function at optimum efficiency. With the Marsona you can simulate the natural sounds of ocean surf, summer rain, mountain waterfalls. You control the volume, the wave pattern, wave or rain rhythm, and the seeming nearness or distance of the source. The 5 inch speaker brings you sounds as close to nature as you will find. UL listed. **$149.95** #2200.

THE LIFESTYLE RESOURCE™

ALLERGY AND COLD BUSTER

Now a major scientific breakthrough — The Viralizer® System—gives you relief from cold, sinus and allergy symptoms. It's the newest development of a concept pioneered at the Pasteur Institute in Paris. The cause of the common cold is the Rhinovirus family which lives and multiplies in the nose and throat, but cannot thrive in temperatures over 110°F. The Viralizer is designed to deliver a gentle, controlled heat which penetrates the nose and throat, creating a hostile environment for cold viruses. After a pleasant heat treatment, the Viralizer dispenses either of two mild, over-the-counter, medicated sprays. Vira-Spray I is an analgesic, anti-bacterial spray. Vira-Spray II is a decongestant that provides temporary relief of nasal congestion due to colds, hayfever, sinusitis or allergies. These therapeutic sprays further discourage the stubborn cold and sinusitis germs so you're less likely to be re-infected or spread your cold to others. The Viralizer can produce effective relief by using it for only 3 or 4 minutes, several times a day. Proven in clinical tests 90% effective in eliminating the symptoms of upper respiratory infection in 24 hours or less, the Viralizer works without pills. Viralizer, is safe for children and adults, has been tested and recommended by doctors. The complete Viralizer® System includes 1 electric Viralizer with Vira-Spray I and Vira-Spray II plus a 3-pak refill of medicated sprays. **$39.95** #1690.

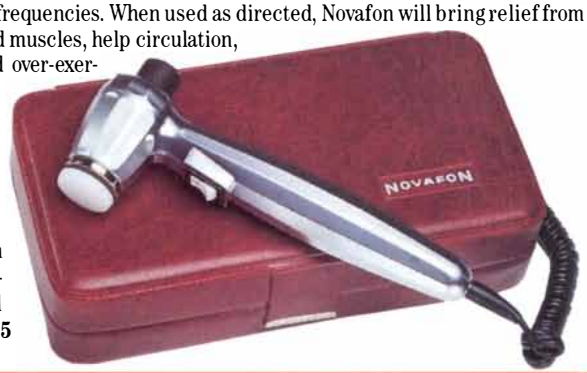
WHEN YOU'RE IN A HURRY

When you're running short of time, ask one of our phone operators about UPS Blue Label delivery. In most cases, we can get your shipment to you in three or four working days.

CALL TOLL-FREE 800-872-5200

SOOTHING SOUND MASSAGE

Tested more than 30 years, the Novafon sound massager is used in clinics and health centers around the world. Novafon's sound waves penetrate up to 2¼", and they help a variety of conditions because of their mixed frequencies. When used as directed, Novafon will bring relief from pain, loosen joints and soothe tired muscles, help circulation, speed recovery from exercise and over-exertion. The Novafon is lightweight (8 oz.) small (8"length). Adjustable intensity control, choice of massage heads (disc-type and ball-type). It comes in a fitted plush case, perfect for carry along. 1 year warranty. A precision made instrument with no interacting parts to wear out, the unit will give many years of service. **$169.95** #1750.



WORTH CUTTING THE CORD FOR

Southwestern Bell FF-1700 Offers—

- *1000-foot range*
- *Base/handset intercom*
- *Memory dialing*
- *Hearing aid compatible*
- *10-channel select*
- *Digital security code*
- *Hold button*
- *Tone/pulse switchable*
- *Jack for answering machine connection*

A leading consumer magazine writer likens a person's first conversation walking around talking on a cordless telephone to the exhilaration of that first bike ride minus training wheels. Later, that article rates Southwestern Bell's FF-1700 Cordless Phone tops for range in controlled tests among 21 brands and models. By handling incoming and outgoing calls to range of 1000′ (the article rated a maximum of 1500 ft.), with outstanding speech quality and convenience features, the FF 1700 ended on top of the consumer magazine ratings reports. Base unit serves as freestanding speakerphone with dialpad, so you get two phones in one. Plus intercom, paging and 10-channel selection. Digital security code protects line from outside access. **$179.95** #2130.

To take your freedom a step further, Southwestern Bell's FA-450 Telephone Answering Machine gives you the latest technology and newest features at a most attractive price. Single cassette operation, call screening, household memo function, voice-activated record, one-touch playback. Two-way record for messages or conversations. Beeperless remote lets you retrieve messages from any pushbutton phone, also allows remote announcement changes. These new-generation Freedom Phones connect you in without tying you down. **$99.95** #2140.

THE LIFESTYLE RESOURCE™

CREDIT CARD ORDERS CALL TOLL-FREE 24 HOURS 800-872-5200

DEPT. SFAH08; 921 EASTWIND DR. SUITE 114: WESTERVILLE, OH 43081

ITEM NO.	QTY.	DESCRIPTION	ITEM PRICE	TOTAL PRICE
			SUB TOTAL	
			SHIPPING (see table at left)	
			TOTAL	

SEND TO (PLEASE PRINT)

ADDRESS

CITY STATE ZIP

☐ CHECK OR MONEY ORDER ☐ MASTERCARD ☐ VISA ☐ AMEX SIGNATURE

ACCT. # EXP. DATE

Shipping Charge covers UPS. handling and insurance for guaranteed delivery.

Up to	$20....	$ 3.95		$50.01	to $60....	$ 7.95	
$20.01	to $30....	$ 4.95		$60.01	to $70....	$ 8.95	
$30.01	to $40....	$ 5.95		$70.01	to $100....	$10.95	
$40.01	to $50....	$ 6.95		Over	$100....	$12.95	

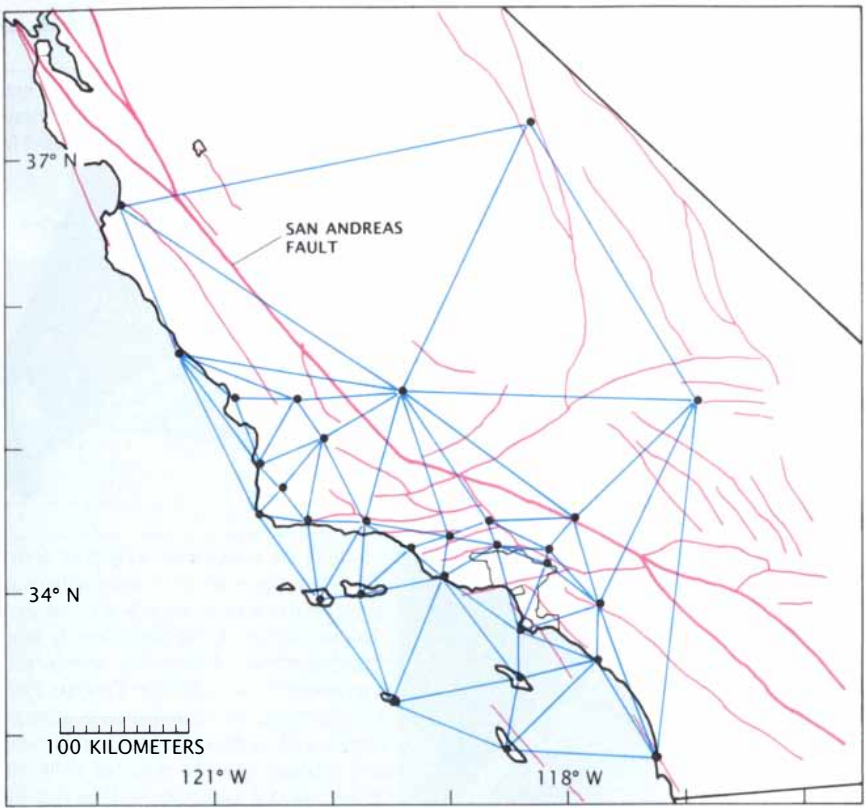
UPS 2ND DAY AIR

UPS Second Day available for an additional $7.50 per order.

ORDER WITH CONFIDENCE

- *We ship most orders within 48 hours.*
- *Credit Card orders billed only upon shipment.*
- *No risk 30-day return privilege.*

THE MEDIA DEVELOPMENT GROUP



NEW GEODETIC NETWORK monitors deformation west of the San Andreas. Positions of stations can be determined to a precision of about one centimeter by means of signals from satellites of the U.S. Department of Defense's Global Positioning System. The network makes it possible to observe changes in the positions of the stations as the crust deforms. It was set up in 1985 by the California Institute of Technology, the Massachusetts Institute of Technology, the University of California at Los Angeles, the University of California at San Diego and several Government agencies.

the rugged coast ranges. Its effects can also be seen offshore, where seismic exploration for oil has revealed a number of thrusts and folds in the rapidly accumulating sediments on the continental shelf. In addition, seismological and morphological evidence suggests that some of the compression actually occurs on structures just east of the San Andreas Fault; for example, the Coalinga earthquake of May 23, 1983, was centered on a point about 27 kilometers east of the San Andreas. It appears likely, therefore, that the compressional motion involves deformation distributed across a region extending from the western edge of the Great Valley to the continental shelf—a region some 150 kilometers wide.

The shear motion, on the other hand, may be much more localized. Much of it may be accommodated, for example, along the San Gregorio-Hosgri fault system, which runs along the coastline from San Francisco to Vandenberg Air Force Base. This fault system is largely submerged, but where it comes onshore there is good evidence that significant displacements have occurred along it during the past few million years. Geological evidence suggests the average rate of motion along the fault may have been as high as 13 millimeters per year. The historically recorded seismicity of this fault system has been very low; if the rate of motion is really as large as 13 millimeters per year, the shear motion would have to be expressed in infrequent but large earthquakes, which are not yet evident in the historical record. It is not clear whether the more modest rate of six millimeters per year deduced from our vector calculations would require such events. The matter is of more than academic interest, however: the Diablo Canyon nuclear power plant and the Vandenberg Space Shuttle Launch Facility are both near the Hosgri Fault, as are a number of large population centers.

These arguments concerning the deformation of California are largely indirect, incorporating data acquired well outside the region. Because we must make many assumptions and simplifications in order to estimate the rate and direction of deformation, our results have large uncertainties and may be biased. If we are to develop more accurate estimates, we must have direct observations of the relative motion of sections of crust on opposite sides of particular fault zones.

To make such observations possible, a consortium of four universities is cooperating with several U.S. Government agencies to set up and monitor a geodetic network that will be distributed across the California margin from Monterey to San Diego and will include most of the offshore islands of the Southern California Borderlands. The network takes advantage of a newly developed technique of space geodesy, one that relies on a constellation of earth-orbiting satellites called the Global Positioning System, which was launched and is being maintained by the U.S. Department of Defense.

The system is designed to pinpoint ships, aircraft and land vehicles in real time with an accuracy of several meters, but methodology borrowed in part from VLBI allows measurements between fixed stations to be made with an accuracy of a few centimeters or less over baselines as long as several hundred kilometers. Hence the system is nearly as accurate as VLBI; moreover, it does not require ground stations as large or expensive as VLBI radio telescopes. Repeated surveys over the next several years should enable geophysicists to develop much more detailed and accurate descriptions of deformation in coastal California. Those descriptions in turn will make possible much more realistic estimates of the risk of earthquakes in this seismically active and highly populated region, and will help us to understand in much greater detail one component of the deformation of the western United States.

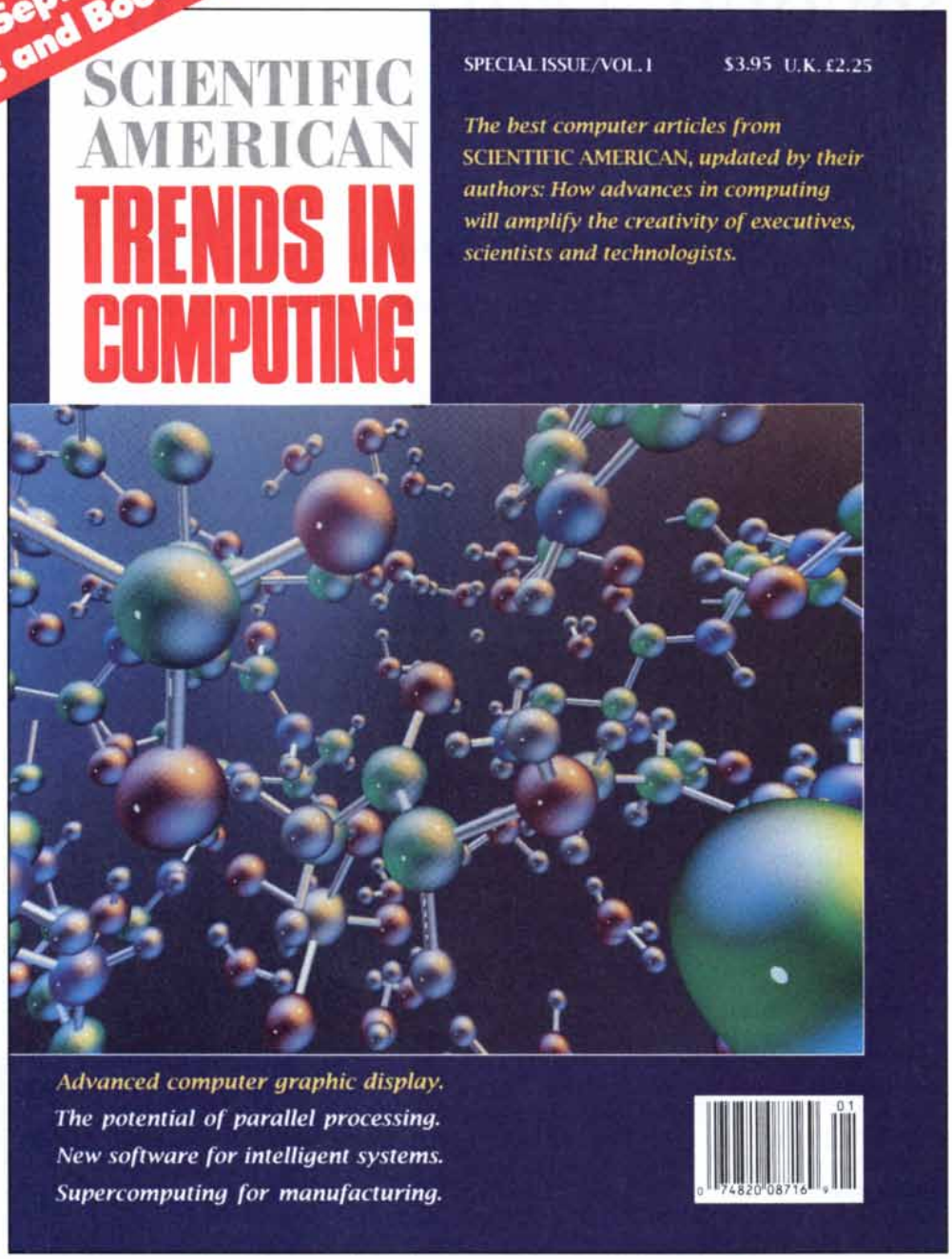
FURTHER READING

CONTEMPORARY BLOCK TECTONICS: CALIFORNIA AND NEVADA. David P. Hill in *Journal of Geophysical Research,* Vol. 87, No. B7, pages 5433–5450; July 10, 1982.

SPECIAL ISSUE ON SATELLITE GEODYNAMICS. In *IEEE Transactions on Geoscience and Remote Sensing,* Vol. GE-23, No. 4, pages 355–552; July, 1985.

VECTOR CONSTRAINTS ON WESTERN U.S. DEFORMATION FROM SPACE GEODESY, NEOTECTONICS, AND PLATE MOTIONS. J. Bernard Minster and Thomas H. Jordan in *Journal of Geophysical Research,* Vol. 92, No. B6, pages 4798–4808; May 10, 1987.

Coming in September to Newsstands and Bookstores



Trends in Computing™, from Scientific American, the most interesting and useful recent articles on the growing impact of computers/computing on American business and science.

At your newsstand or bookstore.

To order direct from the publisher, please complete and return the form, with appropriate remittance to:
Publisher, Scientific American **Trends in Computing™**, 415 Madison Avenue, New York, NY 10017

ORDER FORM

As soon as it comes off the press in September, send me ____ copies of **Trends in Computing™** at US $3.95 each (10 or more: $3.45 each).

I enclose my check/money order (payable to Scientific American) for $_____
(Add $1.00 per copy for orders outside U.S.)

Payment must accompany order.

Name _____

Address _____

City/State/Zip _____

Beyond Truth and Beauty: A Fourth Family of Particles

Three families of the fundamental particles called quarks and leptons are known. Recent experiments hint that there is one more family, but there are probably no more than five

by David B. Cline

Physicists who study the fundamental nature of matter have a faith that the diversity of the physical universe can be explained by assuming the existence of a few fundamental particles. It is a faith that has been sorely tried. In the middle years of this century the emerging simplicity of the proton, neutron and electron and their antimatter counterparts dissolved into hundreds of subnuclear particles. In the 1970's simplicity seemed to reemerge with the discovery of the quark, only to apparently unravel again as several other quarks appeared.

Now the tide of battle may be turning toward a compromise agreement. On the one hand, observations of the isotopes deuterium and helium in deep space, coupled with laboratory accelerator experiments, now indicate that the number of fundamental particles is indeed limited. On the other hand, there are some hints that this number may include more than the three families of quarks now known to exist. Stoking the excitement is the prospect that answers to profound questions such as the origin of mass may be within the reach of sophisticated new accelerators that are just beginning to operate.

DAVID B. CLINE is professor of both physics and astronomy at the University of California, Los Angeles, and a research physicist at CERN, the European laboratory for particle physics. His undergraduate degree is from Kansas State University and his Ph.D. (1965) is from the University of Wisconsin at Madison. He was professor of physics at Wisconsin and a research physicist at the Fermi National Accelerator Laboratory until he took up his joint appointment at U.C.L.A. in 1986. Cline's current interests include the design of a *B*-meson "factory," the construction of a high-resolution gamma-ray telescope and colliding-beam research at CERN.

To understand why some physicists think a fourth family of quarks may exist, but that there are not many more than four, one must first understand what is currently explained and unexplained by the standard model of particle physics. Almost every field has its standard model; the standard model in particle physics is based on the assumption that ordinary matter is composed of two types of particles, quarks and leptons, and that the forces between them are transmitted by a third category of particles called bosons. Leptons include the familiar electron and neutrino; the less familiar quarks combine to make up such large particles as the proton and the neutron. An example of a boson is the common photon, which transmits the electromagnetic force.

Three families of quarks have been discovered experimentally, each consisting of two particles, making a total of six quarks. The first family consists of the "up" quark and the "down" quark. The up quark has a mass of approximately four million electron volts (MeV), or about 1/250th the mass of the proton (which is close to a billion electron volts, or 1 GeV). The mass of the down quark is slightly greater—about 7 MeV. The second family consists of the "strange" quark and the "charm" quark, with masses of about 150 MeV and about 1,300 MeV respectively. The third family consists of the "bottom" quark, known in civilized parts of the world as "beauty," with a mass of 5.5 GeV, and the "top," or "truth," quark—which has yet to be discovered [*see bottom illustration on page 62*].

Whereas a proton has one unit of positive electric charge, quarks have fractional charge. An up quark has a fractional charge of 2/3 and a down quark has a charge of −1/3. A proton consists of two ups and a down, giving the required total charge of 1. The neutron is composed of two downs and an up for a total electric charge of zero. In a similar manner the various quarks can be combined to form all the other known particles that are not leptons or bosons.

Each family of quarks is roughly 10 times as massive as the preceding family. This fact suggests that any new quarks will be very massive. Indeed, recent experiments at CERN, the European laboratory for particle physics, set a lower limit of about 50 GeV for the mass of the undiscovered truth quark. In each family, however, the quark masses lie within an order of magnitude of each other, and so physicists expect the truth quark to have a mass not more than 10 times the mass of the beauty quark. (If future accelerator experiments produce a very massive truth quark, theorists will begin to scratch their heads.)

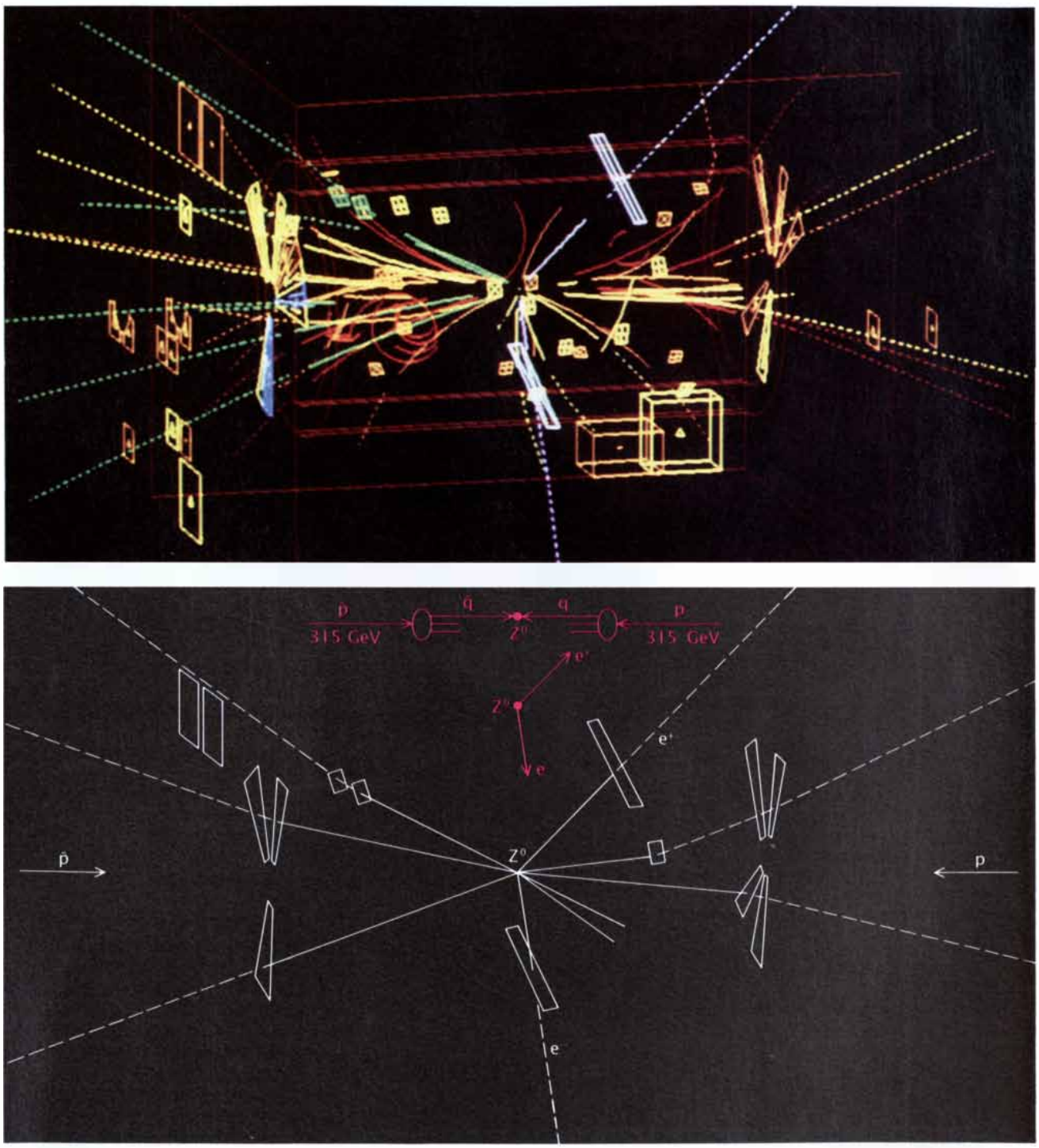
The Lepton Families

Experimentally, it turns out that every family of quarks is associated with a family of leptons, each consisting of a charged lepton and a neutral one. In the first family the charged lepton is the electron and the neutral one is the electron neutrino; in the second family the leptons are the muon and the mu neutrino, and in the third family they are the tauon and the tau neutrino.

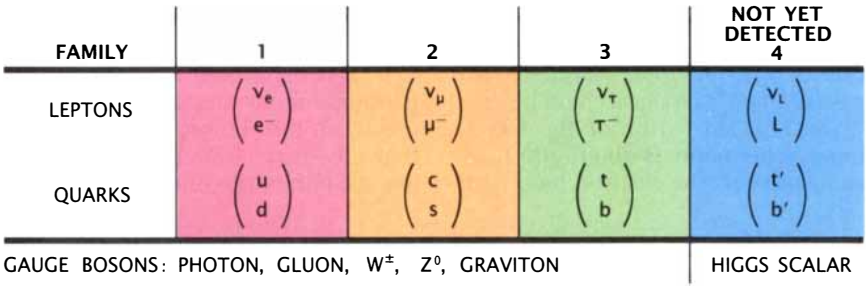
Compared with quarks, leptons are extremely light [*see illustration on page 63*]. The electron has a mass of approximately 1/2 MeV; observations of electron neutrinos from the supernova 1987A limits their mass to less than about 16 electron volts. The rea-

soning is that the velocity of particles having a nonzero rest mass varies with their energy, so that one would expect the arrival time of a burst of massive neutrinos to be spread over a finite period. The fact that the 1987 supernova neutrinos all arrived at the earth within 13 seconds of one another results in the 16 eV limit. And since this is only an upper limit, the real electron-neutrino mass could be zero.

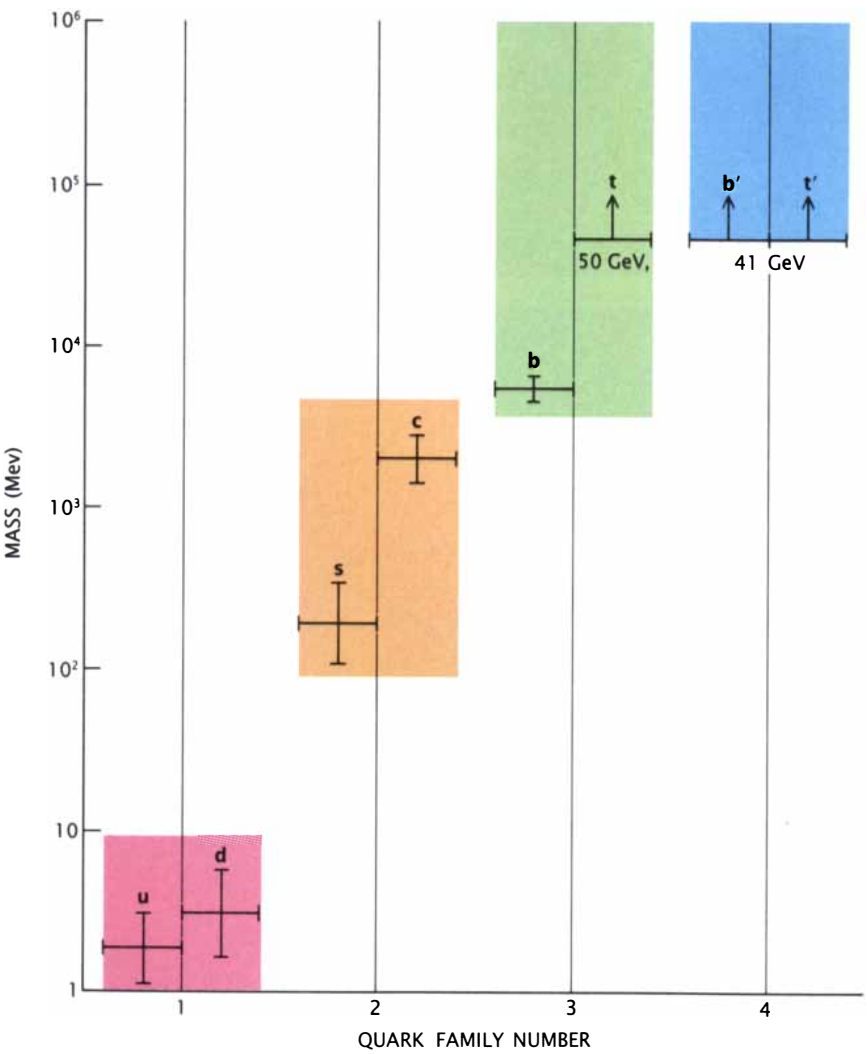
The charged lepton of the second family, the muon, is about 200 times as massive as the electron but is otherwise identical. Experimental limits on the mu neutrino's mass require it to be less than about 100,000 eV. Cosmological limits are much more stringent than this, however; they require that the mass of any neutrino be below the minuscule value of 65 eV. The



PRODUCTION AND DECAY of a neutral intermediate vector boson (Z^0 particle) are shown in a computer image, along with a drawing identifying the events. The Z^0 was produced by quark-antiquark collisions in the UA1 experiment at the proton-antiproton collider at CERN, the European laboratory for particle physics; the particle decays into an electron-positron pair. (Other tracks are those of other particles formed by the colliding beams.) The Z^0 can decay into all existing lepton families, including neutrinos. If there were an infinite number of neutrino families, the decay of the Z^0 into the electron-positron pair would never be seen. The fact that it is seen, with high probability, limits the number of neutrino families to five.



STANDARD MODEL of particle physics assumes that matter is composed of quarks and leptons and that forces are transmitted by bosons. Each lepton family consists of a charged lepton and a much lighter neutral one, successively the electron (e^-) and electron neutrino (ν_e), the muon (μ) and mu neutrino (ν_μ) and the tauon (τ) and tau neutrino (ν_τ). According to the standard model, each lepton family is associated with a quark family consisting of two particles: an "up" (*u*) and a "down" (*d*) quark; a "strange" (*s*) and a "charm" (*c*) quark; a "bottom" (*b*) quark and a "top" or "truth" (*t*) quark. The truth quark has not yet been detected. If the pairing of the families continues to a fourth generation, one would expect to find a fourth charged lepton (*L*) and a fourth neutrino (ν_L), as well as two more quarks, labeled here *t′* and *b′*.



MASSES of the five known quarks are given in millions of electron volts (MeV). The masses of the quarks in successive families differ by about an order of magnitude. Experimental limits on the truth quark put its mass above 50 GeV (billions of electron volts); limits on the fourth-family quarks put their masses above 41 GeV.

tau neutrino has not yet been directly observed, but its partner the tauon was discovered in 1976 by Martin L. Perl and his colleagues at the Stanford Linear Accelerator Center (SLAC). The particle has a mass of 1.8 GeV. Since theory requires that the tauon have its own neutrino, physicists are confident that the tau neutrino exists. Direct experiments put the mass of the tau neutrino at less than 70 MeV; again the cosmological limit appears to be below 65 eV.

Like the quark families, the lepton families are clumped into different mass ranges. The mass of the muon is approximately two orders of magnitude greater than that of the electron, and the tauon in turn is about 20 times more massive than the muon. One might expect, then, that any additional charged leptons will have a mass in the vicinity of 40 GeV. As I shall show below, current experimental lower limits for new charged lepton masses are consistent with this prediction. The neutrino masses may also be spaced at large intervals, but because only upper limits are established, all that can now be said is that the mass of neutrinos is very small compared with the mass of their charged partners.

The smallness of the neutrino mass points to the second major difference between quarks and leptons: apart from the fact that leptons are much lighter in absolute terms than their associated quarks, the mass ratios within quark families are much smaller than those within the associated lepton families. Within each quark family the ratio of quark masses is no greater than about 10. The mass ratio of the down quark to the up quark, for example, is about two. The leptons present quite a different picture: given the upper limits on neutrino masses, the mass ratio of the electron to the electron neutrino is about 10,000; if the neutrino were to turn out to be massless, the ratio would be infinite.

The standard model has also been successful in describing the bosons, the particles that transmit forces between other particles. In Maxwell's theory of electromagnetism this role is played by the photon, which transmits the electromagnetic force. The present standard model contains the weak force, which governs radioactive decay, and the strong force, which binds the nucleus. Therefore other bosons are required. Weak interactions (such as the decay of a neutron into a proton and an electron) that involve the exchange of electric charge are governed by the so-called charged

intermediate vector boson, or *W*. Other weak interactions, which do not require an exchange of charge, are mediated by the neutral intermediate vector boson, or Z^0. One of the great triumphs of the standard model was the prediction of the masses of the *W* and the Z^0 particles. Both particles were subsequently discovered at CERN in 1983—at the expected masses. All other observed properties of the particles are also in remarkable agreement with the theory.

Defects in the Model

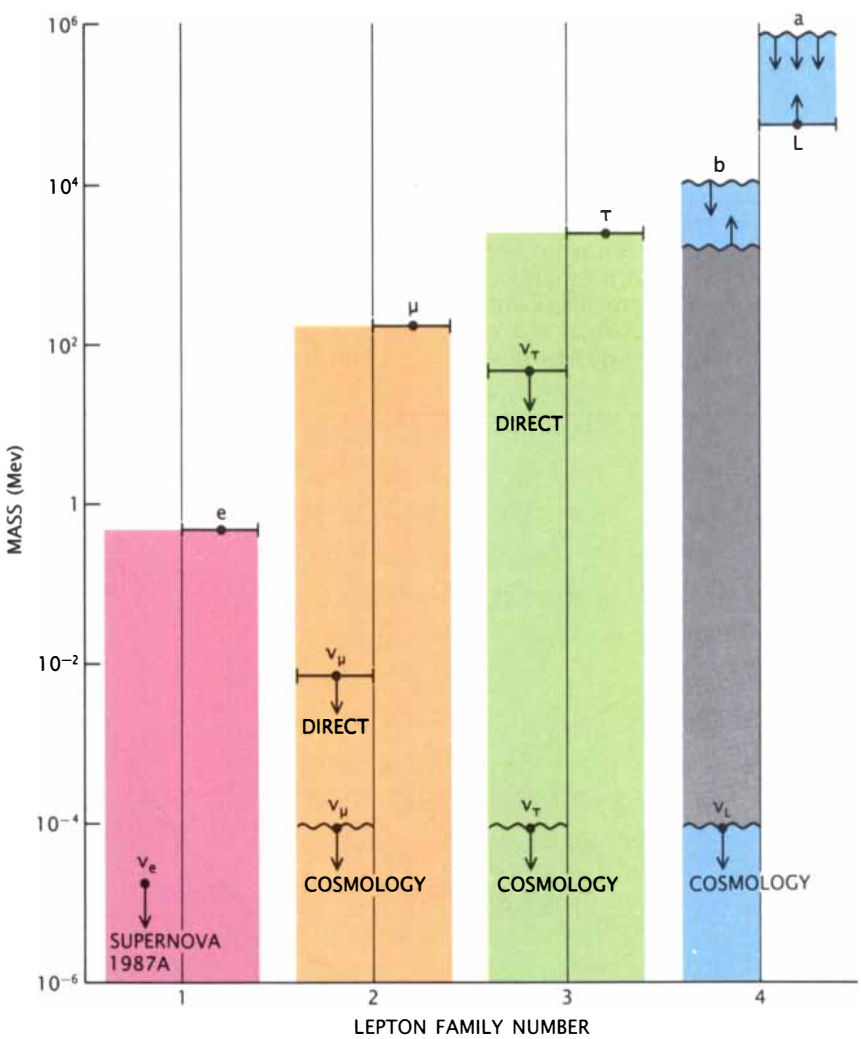
In spite of such successes, the standard model has a number of serious defects. To begin with, it does not prescribe the number of families of quarks and leptons at all. Why are there at least three families, given that only the first family is needed to make up the ordinary protons, neutrons and electrons in the universe? Or, as I. I. Rabi put it 50 years ago, "The muon, who ordered that?"

The standard model also fails to predict the mass of all the remaining particles; the 50-GeV lower limit on the mass of the truth quark is an experimental result, and no one knows what the upper limit is. Nor does the model explain the hierarchy of quark and lepton masses described above. Why are the families separated by roughly an order of magnitude in mass for quarks and two orders of magnitude for leptons? Why are the ratios of quark masses within a family so small and the ratios of lepton masses so large? Many numerological attempts have been made over the years to explain this mass spectrum, but none has met with any success, and this is one of the great unexplained whys of the standard model.

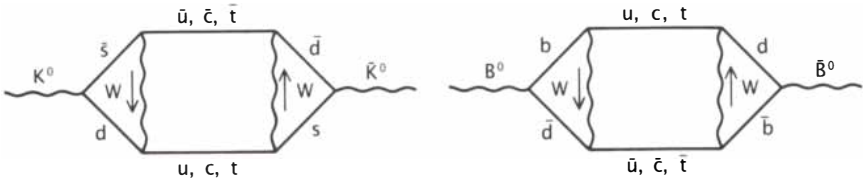
Questions relating to the number of families, mass and mass hierarchies are not the only ones left unanswered by the standard model. Another major mystery is the fact that whereas different kinds of quarks are often observed to transform into one another, leptons are never observed to do so. For example, the charm quark may decay into a strange quark and a particle known as a "virtual *W*" (which can be thought of as a real *W* with such a short lifetime that it cannot be directly observed). On the other hand, no muon has ever been seen to decay into an electron and a photon, and the probability of its happening has now been experimentally reduced to less than one part in 100 billion.

This proliferation of mysteries has led some theorists to suspect the existence of a fourth family of quarks and leptons. The discovery of such a family might clear up some long-standing questions. One of them has to do with the phenomenon known as charge-parity violation, which is itself related to the quark-quark transition probabilities just discussed. Until the 1960's physicists had assumed that measurable properties of any physical system should remain unchanged when each particle is transformed into its antiparticle and the system is reflected in a mirror. Because the operation of changing a particle into its antiparticle requires changing its charge, and because mirror-reflection is known as parity reversal, the statement that any system should remain unaltered under these combined operations is known as the law of charge-parity invariance.

The cherished belief in CP invariance fell in 1964, when Val L. Fitch, James W. Cronin, James H. Christenson and René Turlay of Princeton University were investigating the decay rate of a particle known as the neutral *K* meson, or kaon. The kaon usually decays into three other particles (into three pions, for example); such a transition is consistent with the law of CP invariance. The Princeton experiment showed, however, that about once in every 500 times the kaon decays into only two pions—a transition that vio-



LEPTON MASS SPECTRUM shows that the masses of charged leptons, like those of quarks, are hierarchical: the tauon mass is roughly an order of magnitude greater than the muon mass, which in turn is roughly two orders greater than the electron mass. Limits on the mass of the fourth charged lepton (*L*) require that it be greater than 41 GeV. The mass difference between the *W* and Z^0 vector bosons supplies an upper limit (*a*). The masses of the three neutrinos are not known. Upper limits from supernova 1987A put the electron-neutrino mass at less than 16 electron volts. The requirement that no neutrino be so massive as to noticeably decelerate the expansion of the universe sets a 65-eV "cosmology" upper limit for all neutrinos. The gray area (*right*) is excluded by theory; limit *b* comes from dark-matter searches.



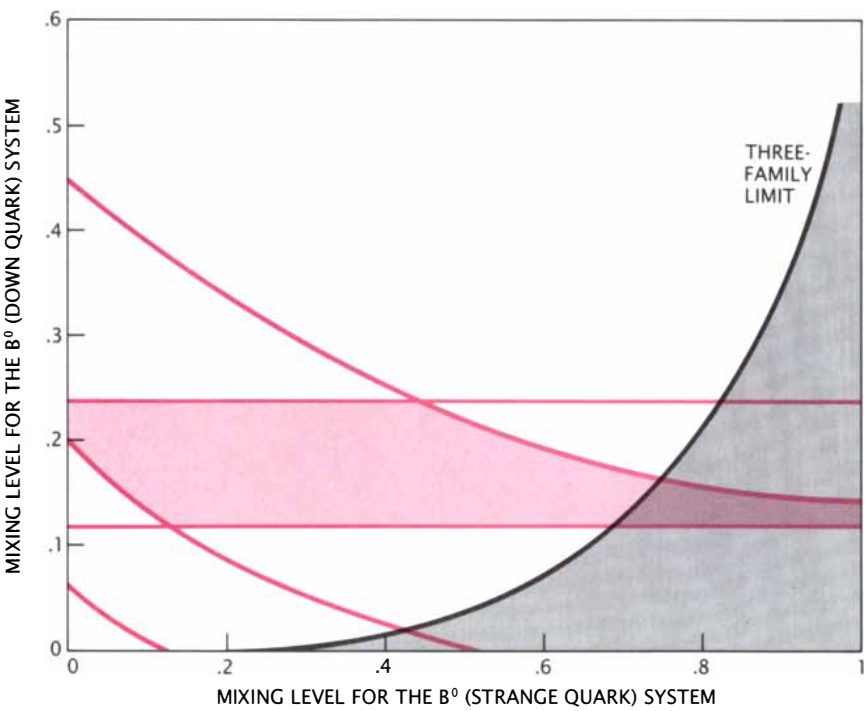
FEYNMAN DIAGRAMS show flavor mixing for the kaon (K^0) system (*left*) and a *B*-meson system (*right*). Mixing takes place when a K^0, which consists of an antistrange ($\bar{s}$) quark and a down (d) quark, turns into an antikaon ($\bar{K}^0$). This requires that the $\bar{s}$ turn into a $\bar{d}$ and the d into an s. In this *B*-meson system, consisting of a b and a $\bar{d}$, the b must transform into a d and the $\bar{d}$ into a $\bar{b}$. The transitions are called cross-family or flavor-mixing transitions. Such mixing is needed for charge-parity violation.

lates CP invariance. As a result of experiment, then, a supposed law of nature was thrown out—even though the actual origin of CP violation in the kaon system remains unexplained today and is considered one of the great mysteries in physics.

Flavor Mixing

Unexplained though CP violation may be, its magnitude can be linked to quark-quark transition probabilities. Usually quarks transform into other members of their own family, as in the decay of a charm quark into a strange quark and a virtual *W*. For CP violation to take place, quarks must be able to transform into members of other families, a process known as flavor mixing (because quark families are whimsically characterized as having distinct flavors). Moreover, it can be shown that a two-family standard model would not be enough to allow CP violation in the neutral kaon system; at least three families of quarks are necessary. Indeed, the existence of CP violation was the first evidence for a third quark family. The amount of CP violation depends on the probability with which a quark from one family can transform into a quark from another family, that is, on the extent of the mixing.

A fourth family would influence the amount of flavor mixing by allowing more quark-quark transitions. Recently physicists at the Deutsches Elektronen-Synchrotron (DESY) and CERN have found that flavor mixing in the *B*-meson system is 20 times as large as expected. The *B* meson is so called because it consists of a beauty quark and one other quark—for example, an antidown quark. In the mixing process the *B* meson is transformed into an anti-*B*, which requires that the beauty quark be transformed into a down and the antidown into an antibeauty [*see top illustration on this page*]. Note that, as in the kaon system, these are flavor-mixing transitions. The rate at which the mixing occurs depends on all existing quarks, as well as on their masses: the more quarks there are, the more mixing is expected to take place. The fact that the mixing discovered at DESY was much larger than expected may indicate that a fourth family of quarks was contributing to it. Because the mass of the truth quark is not yet known, however, it may be that the results can still be accommodated with three quark families.

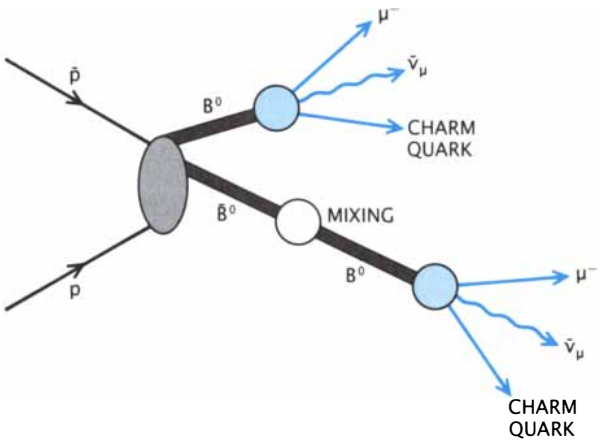


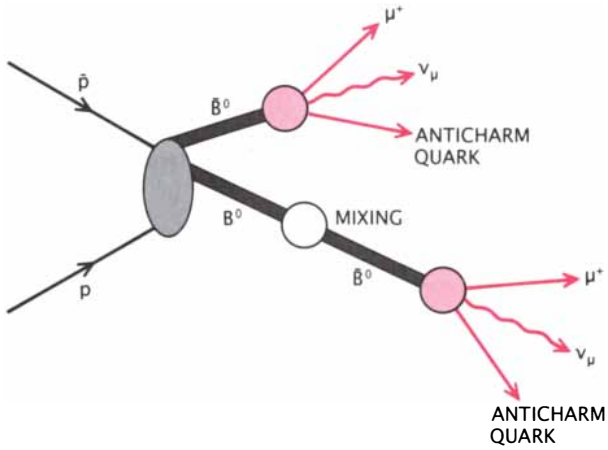
MIXING EXPERIMENTS may provide evidence for a fourth family of quarks. B^0 mesons come in two varieties, either a $b\bar{d}$, as in the preceding illustration, or a $b\bar{s}$. Both types undergo flavor mixing, that is, they change into their antiparticles. Recently five experiments (*colored curves*), done at CERN, DESY, SLAC and Cornell University, have shown the mixing was much larger than expected. The colored area delimited by the experimental results indicates the most probable values of the mixing level for the two types of B^0 particles. Most of that area lies beyond the limit allowed by a three-family standard model, indicating that a fourth family may contribute to the mixing.

B-meson mixing could also provide insight into the origin of CP violation itself, which until now has been observed only in the neutral kaon system. Flavor mixing is a necessary condition for CP violation, but it is not sufficient. Although CP violation has not yet been detected in *B* mesons, the very size of the mixing has made some investigators optimistic that observation of CP violation in the *B*-meson system cannot be far behind. If CP violation is found to be similarly large, it is unlikely that the three-family standard model will be able to accommodate it (unless the mass of the truth quark is unexpectedly large), and so a fourth family of quarks will have to be invoked.

An experimental test of that proposition should be possible in the near future. Proton-proton collisions in an accelerator can produce *B*-meson–*B*-antimeson pairs, which in turn will decay into products containing two charged leptons. These charged leptons might be electrons or positrons. If CP is conserved, the rates at which mesons decay into electrons and into positrons should be the same; if CP is violated, the decay rates will differ. This test will be extremely sensitive to the existence of a fourth quark family. An observation of CP violation would uniquely relate CP violation to quark transition probabilities and could lead to identification of the fundamental origin of CP violation in nature.

The importance of detecting charge-



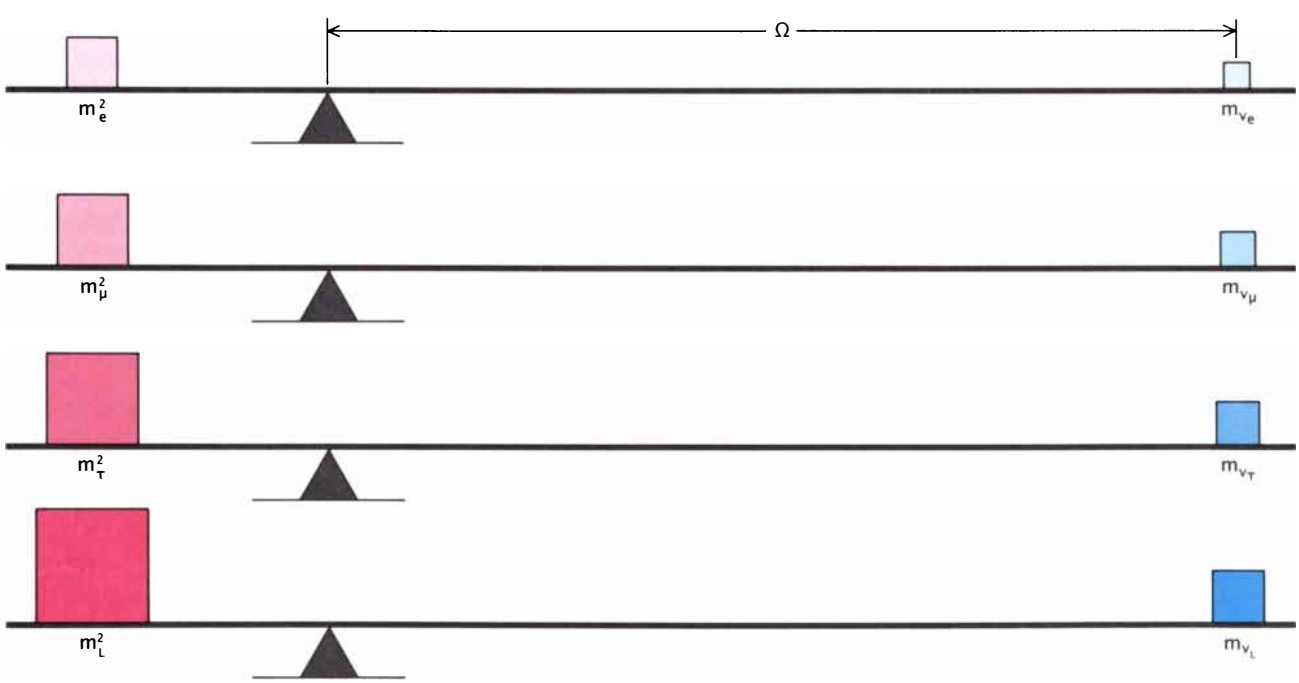


***B*-MESON DECAY** may exhibit charge-parity violation. Proton-proton collisions in an accelerator produce meson-antimeson pairs. A B^0 may decay through flavor mixing into a $\bar{B}^0$, as is shown at the left, or vice versa, as is shown at the right. These particles thereupon decay further into muons, antimuons, quarks and neutrinos. If the decay rate into muons (*left*) differs from the decay rate into antimuons (*right*), then CP is violated. (Note: Charge is conserved here; add the diagrams.)

parity violation in the *B*-meson decay system is leading to the design of a new type of electron-positron collider called a linear-collider *B*-meson factory. Studies of such a machine are in progress at the University of California at Los Angeles and in Italy. The goal is to produce more than a billion beauty quarks and antiquarks per year.

The other major mystery that might be solved by invoking a fourth family is the origin of the particle-mass hierarchy. The hope is that the fourth family is a special case, and that the masses of the first three families are "generated" by interactions with the fourth. This concept, first described by Harald Fritzsch of the University of Munich, relates the mass difference between quarks to an assumed relation between quark mass and quark transitions. So far all observational data are consistent with the Fritzsch model.

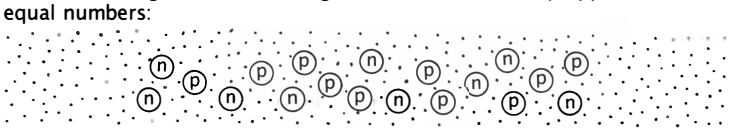
Fritzsch also suggests that the mass ratio of the two new quarks will be four. That is the ratio of the squares of the quark charges—$(2/3)^2/(-1/3)^2$—that would be expected if the electro-



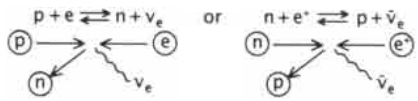
SEESAW MECHANISM proposes that the mass of each neutrino (m_ν) is related to the mass of its associated charged lepton (m) by the formula $m_\nu = m^2/\Omega$; Ω is an unknown mass scale, visualized here as the lever arm of the seesaw. Since, for example, the electron-neutrino mass is known to be less than 16 eV and the electron mass is known to be .5 MeV, the seesaw equation requires Ω to be at least 16 GeV. The tauon mass is 1.8 GeV and cosmological limits on the tau neutrino make it less than 65 eV. Running the seesaw equation with these values gives the stricter lower limit on Ω of 5×10^7 Gev. If Ω is related to a large fourth-lepton mass, the seesaw mechanism shows how this large mass could generate the very small neutrino masses.

BIG-BANG HELIUM PRODUCTION AND NEUTRINO FAMILIES

1. Assume that the universe consists only of neutrons (*n*) and protons (*p*), with a vastly larger background of electrons (e^-), positrons (e^+), neutrinos and antineutrinos ($\nu_e, \nu_\mu, \nu_\tau, \bar{\nu}_e, \bar{\nu}_\mu, \bar{\nu}_\tau$) and photons ($\gamma$), all indicated below by dots. At times much less than one second after the big bang and temperatures much higher than 10^{10} degrees Kelvin, the *n* and *p* appear in almost equal numbers:

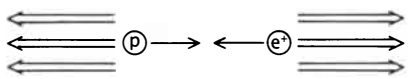


2. Neutrons and protons are constantly transmuted into one another by the so-called weak nuclear reactions:



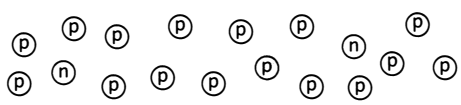
3. Because neutrons are slightly more massive than protons, they are energetically more difficult to produce, and so the *n-p* transmutations in step (2) result in slightly more protons. As the universe expands and cools, less and less energy is available to produce neutrons, and so the weak reactions result in ever more protons. At about one second after the big bang and a temperature of about 10^{10} degrees K. protons outnumber neutrons by about five to one:

4. At this time the expansion rate of the universe overtakes the ever slowing weak-reaction rates, so that collisions between particles essentially cease:

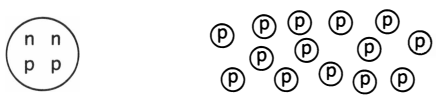


No more neutrons are converted into protons; the 1:5 ratio is "frozen out."

5. Neutrons are radioactive and decay into protons. The lifetime of the neutron is about 15 minutes, so that after three minutes or so about one-third of the neutrons have decayed into protons, leaving one *n* for every eight *p*:



6. At three minutes after the big bang the temperature has dropped to about 10^9 degrees K., which is low enough so that the nucleus of the isotope deuterium (*n,p*) can stay bound. Deuterium is then rapidly processed into helium (2*n*,2*p*). Since helium requires equal numbers of *p* and *n*, helium formation ceases when all the available neutrons are used up:



Since neutrons and protons are of almost equal mass, about 4/16, or 25 percent, of the mass of the universe ends up in helium, with 75 percent left over in protons (hydrogen nuclei).

7. The more families of neutrinos there are, the faster is the expansion rate of the universe. Step (4) therefore occurs earlier and at a higher temperature when more neutrons are present; steps (5) and (6), then, proceed in the presence of more neutrons, resulting in the formation of more helium. Astronomical observations, however, limit helium to less than 25 percent of the mass of the universe. This in turn indicates that there are no more than four neutrino families.

magnetic interaction gives rise to the mass of the quarks.

Several theorists think a new quark should exist in the vicinity of 246 GeV. One of the notable features of the standard model is its prediction that at high enough energies the various forces begin to unify. In particular, the electromagnetic force and the weak and strong nuclear forces should become a single "grand unified" force. The forces should be unified at the incredible energy of 10^{15} GeV, considerably beyond what can ever be attained by an accelerator on the earth. The extrapolation of measured values of fundamental parameters from low energies to the grand-unified energy scale would require the existence of a new massive quark for consistency.

Furthermore, it turns out that the measured values for the W and Z^0 masses can provide limits on the mass difference between the two members of the fourth quark family. Present data show that the mass difference between the two quarks of the fourth family should be less than 180 GeV. Hence if one member of the family does exist in the vicinity of 246 GeV, the mass of the other member should lie either at 426 GeV or at about 66 GeV, the latter being well within the range to be searched by accelerators in the near future.

Neutrino Masses

If the fourth family of quarks exists, what are its lepton relatives like? Here the interest centers on the neutrino masses. The cosmological limits mentioned above require the neutrinos to have a mass less than 65 eV—which includes zero. When compared with the mass of the W, this gives a mass ratio of one billion. What accounts for the incredibly small neutrino mass? There are two different viewpoints: either the neutrino mass is exactly zero as the result of some undiscovered fundamental principle, or the small neutrino mass is a consequence of another very large mass.

The latter viewpoint depends on what is known as the seesaw mechanism, which has been proposed by Murray Gell-Mann of the California Institute of Technology, Pierre M. Ramond of the University of Florida and Richard C. Slansky of the Los Alamos National Laboratory. The disadvantage of the seesaw mechanism is that it is ad hoc; the advantage is that the mechanism is extremely simple. It assumes that the electron-neutrino mass is equal to the square of the electron mass divided by some large

unknown mass scale. The electron mass is fixed. Therefore the larger the unknown mass scale, the smaller the resulting neutrino mass; hence the name seesaw [*see bottom illustration on page 65*].

To illustrate, the supernova limits put the electron-neutrino mass at less than 16 eV. The square of the electron mass is about 250 GeV^2. The solution of the seesaw equation then requires that the unknown mass scale be greater than 16 GeV. When the calculation is run for the 1.8-GeV tau mass and the 65-eV cosmological upper limit on the tau-neutrino mass, one finds a more stringent lower limit of 5×10^7 GeV.

One sees that by the seesaw mechanism the incredibly tiny neutrino mass is the consequence of a mass scale much greater than what is attainable by present colliding-beam accelerators. If the mechanism is correct, these mass scales should be associated with new particles—perhaps a fourth quark or lepton. The seesaw mechanism would then have cosmological implications as well: it raises the possibility that the fourth neutrino might provide the so-called missing mass needed to close the universe.

Current theoretical prejudice requires that the mass density of the universe be just sufficient to eventually halt the present expansion and cause the universe to recollapse, in which case the universe is said to be "closed." The available evidence, however, indicates that the observed mass density of the universe is only between 10 and 20 percent of this critical value. Astronomers are therefore now engaged in an extensive search for the "missing mass."

A neutrino that provides the missing mass cannot be too massive. Neutrinos are even more plentiful than photons—several billion for every proton, electron and neutron—and if any one type of neutrino had a mass equal to the 65-eV value, that would be enough by itself to close the universe. If the neutrino mass were much above this value, the resulting gravitational pull would be sufficient to slow the observed expansion rate of the universe noticeably. The fact that no such effect is observed has led most physicists to accept 65 eV as an upper limit.

Now, it is known from experiment that any fourth charged lepton must have a mass greater than 41 GeV. Given this number for the charged lepton's mass and 65 eV for the fourth neutrino's mass, the seesaw mechanism yields a value of 2.5×10^{10} GeV for the unknown mass scale. Assuming that this single mass scale generates the masses of all the neutrinos, one then computes by the seesaw mechanism that the neutrino masses must be less than 10^{-8} eV, 4×10^{-4} eV and .1 eV respectively for the electron neutrino, mu neutrino and tau neutrino. If this argument is correct, the fourth neutrino could provide the missing mass, but the three neutrinos already known would be much too light to have any effect.

Experiments Under Way

Such arguments for the existence of a fourth family of quarks and leptons are admittedly speculative. Yet some direct searches are under way. One technique was first suggested by the author and Carlo Rubbia of CERN. It uses the decay of the W particles to discover, or to put a limit on, the mass of a possible fourth charged lepton. Recent experiments at CERN give a 41-GeV limit. Notice that this is between one order and two orders of magnitude more massive than the tauon, which is what one would expect from the mass hierarchy discussed above. If the mass of the next quark or charged lepton is less than 70 GeV, present-day machines may be able to detect them in the near future. Otherwise physicists will have to wait for the Superconducting Supercollider or for the Large Hadron Collider that has been proposed at CERN.

One might well wonder whether, if a fourth family of quarks and leptons is uncovered, a fifth will be far behind. This question is being addressed by both cosmologists and particle physicists in ongoing attempts to count neutrino families.

From the cosmological standpoint, the number of neutrino families has a profound effect on the production of light isotopes in the process of primordial nucleosynthesis that occurred in the first few minutes after the big bang. The final abundances of these isotopes, in particular helium and deuterium, depend on how fast the universe was expanding in relation to the rate of the isotope-producing nuclear reactions. The expansion rate of the universe in turn depends on the number of particle species in existence, including families of neutrinos. The more neutrino families there are, the faster the universe expands and the more helium is produced. Comparing the helium produced by nucleosynthesis calculations with observational upper limits then constrains the number of possible neutrino families [*see box on opposite page*]. Remarkably, such considerations limit the number of neutrino families to four, or conceivably five. Assuming that the quark-lepton pairing of the standard model continues to higher families, this then limits the number of quark families to four or five too.

Such cosmological limits on particle species have traditionally been taken with large grains of salt by particle physicists. Now, however, laboratory experiments are coming to the same conclusion. These experiments utilize the Z^0 particle, which is able to decay into all existing neutrino families. The more neutrino families there are, the faster the Z^0 is able to decay. Hence, by measuring the Z^0 lifetime, physicists can determine the number of neutrino families existing in nature. Preliminary results from SLAC, CERN and DESY have already limited the number of neutrino families to five. Refinements will eventually give an exact number—not just an upper limit: direct evidence for or against a fourth family of quarks and leptons.

If a family beyond truth and beauty is established, physicists will no doubt ask: Why four? Why are the fourth-family masses so large? Why is four (an even number) also the number of dimensions of spacetime? Is this the result of superstring theory, which specifies the number of possible dimensions? Admittedly, questions that ask why lead to infinite regress. Yet the only thing science can do is to reduce many problems to a few. If the introduction of the fourth family of quarks and leptons explains the mass distribution of the first three families, transition probabilities, missing mass and the nature of charge-parity violation—then it will have accomplished a great deal.

FURTHER READING

WEAK-INTERACTION MIXING IN THE SIX-QUARK THEORY. H. Fritzsch in *Physics Letters,* Vol. 73B, No. 3, pages 317–322; February 27, 1978.

SEARCH FOR HIGH-MASS SEQUENTIAL LEPTONS THROUGH THE CHANGE OF CHARGE ASYMMETRY IN $W^{\pm} \rightarrow l^{\pm}$ + (NEUTRINOS) DECAY. D. B. Cline and C. Rubbia in *Physics Letters,* Vol. 127B, No. 3, 4, pages 277–280; July 28, 1983.

PRIMORDIAL NUCLEOSYNTHESIS: A CRITICAL COMPARISON OF THEORY AND OBSERVATION. J. Yang, M. S. Turner, G. Steigman, D. N. Schramm and K. A. Olive in *The Astrophysical Journal,* Vol. 281, No. 2, Part 1, pages 493–511; June 15, 1984.

THE FIRST INTERNATIONAL SYMPOSIUM ON THE FOURTH FAMILY OF QUARKS AND LEPTONS. In *Annals of the New York Academy of Sciences,* Vol. 518; 1988.

Light-activated Drugs

A patient ingests an inert substance. The substance is removed from the body in a small amount of blood and activated by light. The result? Effective treatment for a stubborn cancer

by Richard L. Edelson

One of the most highly prized goals of medicine is the development of therapeutic agents with a precise specificity: agents that affect the target tissue and no other. Many antibiotics approximate this ideal. Because of the differences between bacterial cells and human ones, antibiotics can destroy the invader and leave the body's tissues unharmed. In cancer and in autoimmune disorders (where the immune system attacks normal tissues), however, such discrimination is much harder to achieve. In both types of disorders the cause of the tissue damage is not a bacterium or a virus but the body's own cells gone awry. Clearly, distinguishing between diseased cells and healthy ones is no easy task. As a result there are few effective therapies for autoimmune diseases, and those that exist for cancer often carry with them forbidding side effects.

One strategy that has begun to pay significant dividends is the exploitation of drugs that are activated by light. Because these drugs are inert when they are not exposed to the correct wavelength of radiation, they enable the clinician to target only those tissues exposed to both drug and light. In particular, my colleagues and I have made use of a drug known as 8-MOP, whose history extends back to the ancient Egyptians, in treating cutaneous *T*-cell lymphoma (CTCL). CTCL is a malignancy of white blood cells with a poor prognosis. By giving patients 8-MOP, however, and then subjecting their white blood cells to the right wavelength of radiation, we have achieved some striking results. This procedure, now approved by the U.S. Food and Drug Administration for treating CTCL, may presage a wide range of therapies based on light-activated drugs.

RICHARD L. EDELSON is professor in and chairman of the department of dermatology at the Yale University School of Medicine. His research centers on human *T* cells and their malignancies. This is his second article for SCIENTIFIC AMERICAN.

As it happened, our work ultimately brought together two streams of experiments that had initially been widely separated: one on blood-cell cancers and the other (carried out by Irun R. Cohen of the Weizmann Institute of Science in Israel and his associates) on autoimmune disorders. These areas of investigation shared several elements. In the first place, both the cancers my colleagues and I studied and the autoimmune disorders studied by Cohen stem from abnormalities of *T* lymphocytes, white blood cells that have a central role in the orchestration of the immune response. (It is a certain subset of *T* cells that is affected in AIDS.)

What is more, physicians attempting to treat both types of disease frequently encounter analogous problems. Cancer chemotherapy is typically intended to affect rapidly multiplying cells. The patient's tolerance for such therapy is limited by its toxicity to normal tissues—such as the lining of the intestine, the bone marrow and the hair follicle—that include rapidly dividing cells. Those undergoing cancer chemotherapy often suffer from anemia (as a result of decreased production of red blood cells by the bone marrow), bleeding (the result of decreased production of blood platelets, which are required for clotting) and infection (the result of decreased production of white blood cells), as well as intestinal bleeding and hair loss.

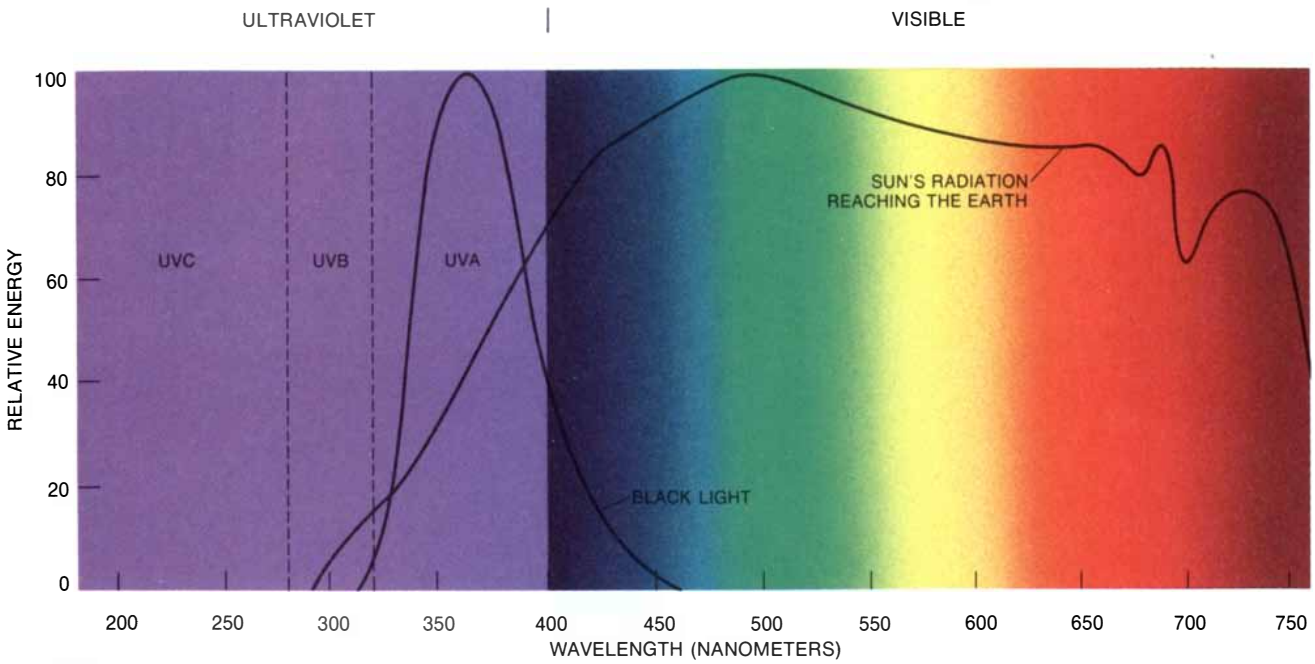
Much as the cancer chemotherapist attempts to interfere selectively with malignant cells, the clinical immunologist tries to stem the growth of abnormally reactive cells of the immune system. In many instances the abnormal cells are *T* lymphocytes. For example, in rheumatoid arthritis, an aggressive and disfiguring disease that mainly strikes young women, the lymphocytes attack joint tissues as if those normal body tissues were an invading pathogen. Cortisone derivatives are often employed to suppress the destructive lymphocytes. Yet among the common side effects of cortisone derivatives are hypertension, diabetes, cataracts, bone degeneration and susceptibility to infections that would be handled with ease by a healthy person.

Although it has now become clear that there is an intriguing overlap between the problems of the immunologist and those of the cancer specialist, in the beginning my colleagues and I were concerned mainly with light and cancer. Light has long been recognized as having a role in the management of disease. In 1903 Niels R. Finsen received a Nobel prize for his finding that skin lesions of tuberculosis often resolved after exposure to ultraviolet light. Tuberculosis is currently treated with antibiotics, but light therapy is central to managing several common disorders, including psoriasis, acne and some forms of jaundice in newborns.

Long before Finsen, however, the ancient Egyptians had recognized that a common plant had medicinal properties that were elicited by light. The plant, *Ammi majus,* is a weed that grows on the banks of the Nile. The physicians of the time noted that soon after eating the plant people became unusually prone to sunburn. They exploited this property to treat the skin disorder vitiligo, in which the skin appears blotchy because some areas lose their pigmentation. It is now known that the active ingredients in the *Ammi* plant are psoralens, the class of compounds to which 8-MOP (its technical name is 8-methoxypsoralen) belongs. Today 8-MOP, which is both an anticancer drug and an immune modulator, is a prototype for the development of drugs that can be activated by light.

HANDS BEFORE TREATMENT are those of a patient with cutaneous *T*-cell lymphoma (CTCL), a cancer originating in white blood cells called *T* lymphocytes. *T* cells are present in large numbers in the skin. Their abnormal multiplication yields the distorted appearance seen here. Ultimately malignant cells invade internal organs. CTCL resists conventional therapies.

HANDS AFTER TREATMENT with a light-activated drug called 8-MOP are the same ones shown in the photograph at the top. 8-MOP is inert until it is exposed to ultraviolet radiation of a particular wavelength. Then it binds to DNA, damaging rapidly multiplying cells such as cancerous *T* cells. The treatment for CTCL, which was devised by the author, is called photopheresis.



ELECTROMAGNETIC SPECTRUM includes both visible and invisible radiation. Among the invisible forms is ultraviolet radiation. Ultraviolet B (UVB) causes sunburn. Ultraviolet A (UVA) includes the wavelengths required to activate 8-MOP.

The psoralens are an extraordinary group of compounds, and it is to them that most of our attention has been directed. The psoralens are found not only in the *Ammi* plant but also (in smaller quantities) in figs, limes, parsnip roots and many other fruits and vegetables. After being taken orally, psoralens are absorbed from the digestive tract, reaching peak levels in the blood and other tissues in from one to four hours; within 24 hours they are excreted almost entirely. The most dramatic property of these substances is their capacity to absorb ultraviolet light and be activated by it. Before exposure to that radiation the molecules are inert, but afterward they bind to DNA in such a way that the strands of the DNA helix are firmly linked. Since separation of the strands is required for DNA replication, the psoralens inflict heavy damage on rapidly dividing cells.

Several features of the psoralens make them well suited for clinical applications. In the absence of light they are extremely safe, yet once activated they severely interrupt DNA function. Furthermore, the type of radiation that activates them, ultraviolet A, or UVA, readily passes through glass and certain plastics, unlike ultraviolet B (the cause of sunburn), which is filtered out by those materials. That feature is of great advantage in equipment design. Perhaps even more significant, once psoralens are exposed to UVA, they remain active for only a few millionths of a second, long enough for a chemical reaction but brief enough so that when the radiation is turned off, the drug reverts immediately to the inert form. In essence the psoralens offer the clinician a highly specific form of therapy: drugs potent only when subject to the appropriate radiation.

Intriguingly, the modern medical history of the psoralens is in a sense continuous with its therapeutic use in ancient Egypt. The first modern studies on 8-MOP were done by Abdel M. El Mofty of the University of Cairo in the 1940's. El Mofty confirmed what his ancient predecessors had known: that ingestion of a preparation from the *Ammi* plant in conjunction with exposure to sunlight was an effective treatment for vitiligo. His results greatly interested a group at the University of Michigan School of Medicine led by Aaron B. Lerner and Thomas B. Fitzpatrick, who carried out the first studies of purified 8-MOP. Their work, done in the early 1950's, showed that the drug was quite safe: the ratio of the maximum safe dosage to that required for clinical effect was very high.

After that initial demonstration 8-MOP remained the focus of some experimental interest. During the 1970's several investigators showed that the drug, in combination with light, could be exploited to treat psoriasis. (Psoriasis is a noncancerous condition in which the cells of the epidermis divide much more quickly than normal, resulting in a thick, scaling skin.) At about the same time Barbara Gilchrest and her colleagues at the Harvard Medical School found that the localized skin lesions of CTCL disappeared in response to a similar regimen.

That finding was of interest, partly because CTCL is the commonest adult malignancy of *T* lymphocytes and was stubbornly resistant to most treatments known in the 1970's. In its initial phase the disease is characterized by skin lesions containing many malignant *T* cells. It might be thought strange that white blood cells are found in the skin, but in fact they are common there, because the skin has important immune functions [see "The Immunologic Function of the Skin," by Richard L. Edelson and Joseph M. Fink; SCIENTIFIC AMERICAN, June, 1985]. In the later phases of CTCL the malignant cells are disseminated throughout the body and the prognosis is bleak. Indeed, average survival for an untreated patient is only about five years from the time of a diagnosis confirmed by biopsy.

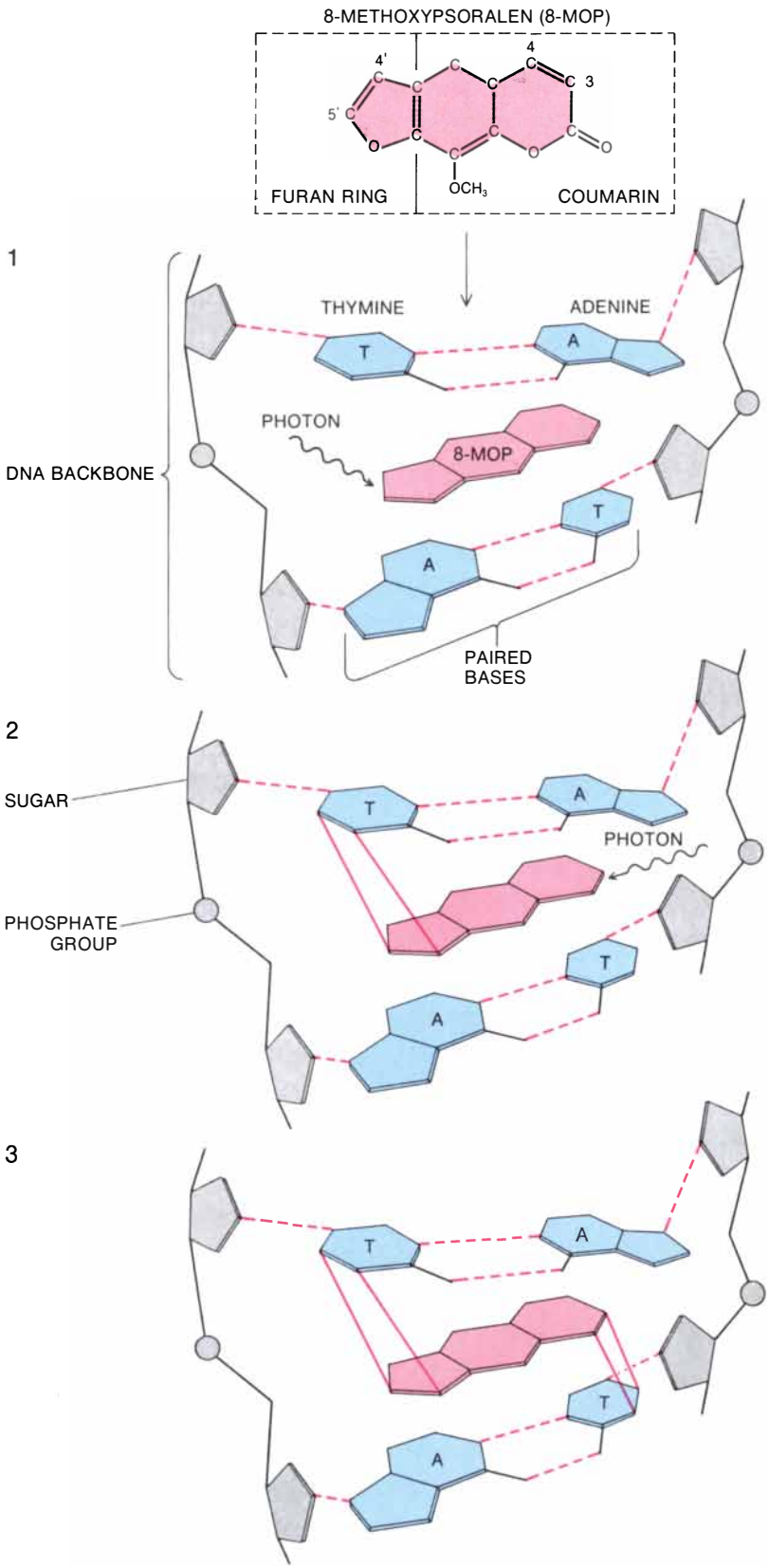
It seemed likely that Gilchrest's experimental combination of 8-MOP and high-intensity, full-body UVA exposure worked because the drug and the light together were directly damaging the diseased *T* cells in the skin lesions. It had long been known that as much as 25 percent of the blood supply passes through the vessels of the skin and that some UVA can pass through the

full thickness of the skin. Moreover, work in my laboratory and in those of Kenneth H. Kraemer at the National Cancer Institute and Warwick L. Morison at Harvard had revealed that *T* lymphocytes are quite sensitive to the damaging effects of 8-MOP and ultraviolet radiation. These findings suggested the tantalizing possibility of removing the malignant cells from the body and subjecting them directly to the toxic pairing of drug and light.

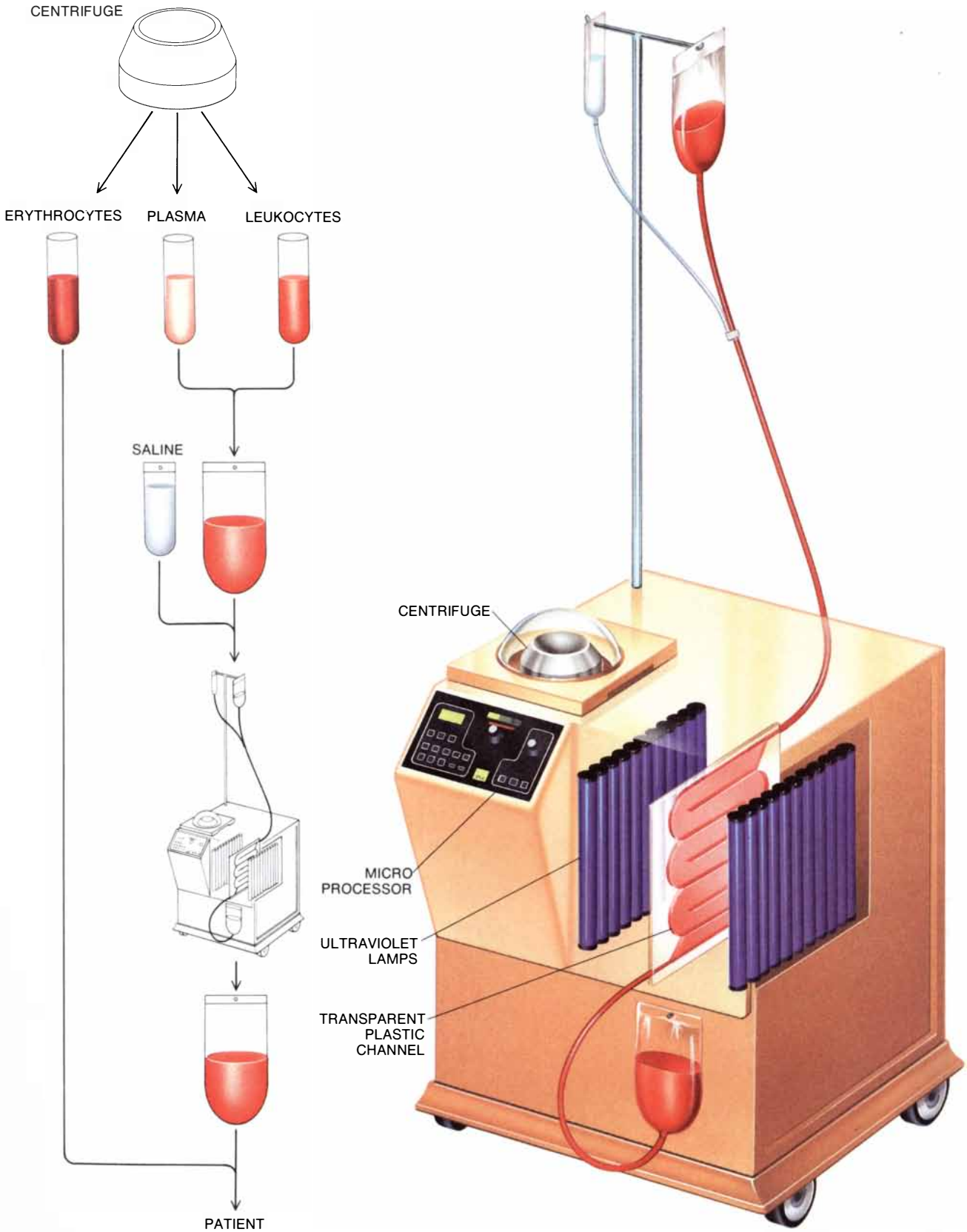
That possibility suggested itself to me partly because I was already working with a system for taking blood from the body and removing the malignant cells from it. Called leukapheresis, the system entails separating the various components of the blood by centrifugation. Because the components have different densities, they segregate in the centrifuge tube and can be parted. The separation, however, is not exact, and as a result some red blood cells, platelets and normal white blood cells are removed along with the diseased lymphocytes. When the procedure was pushed to its limits in an attempt to retrieve the majority of the malignant cells, patients became anemic (owing to red-cell deficiency) and vulnerable to infection (owing to white-cell deficiency).

It seemed the efficiency of this process could be increased if the white cells were exposed to 8-MOP and light once they were out of the body. Since neither red blood cells nor platelets have nuclei, only white blood cells would be affected by the DNA-binding drug. In addition, once it was returned to the body the drug would be inert, and so no tissues other than the blood would be exposed to 8-MOP in its activated state. My colleagues and I looked forward to trying this promising new method. We were somewhat concerned about the consequences of returning large numbers of damaged *T* cells to the patient's bloodstream. What we could not have foreseen was that this would turn out to be the key to the therapy.

After preliminary studies designed to determine the optimum dosage of drug and light, we were ready to proceed. UVA penetrates clear tubing made of acrylic plastic quite efficiently and that material was chosen in the design of the exposure system, a project in which my colleagues and I collaborated with engineers at Therakos Inc., a subsidiary of Johnson & Johnson. UVA is a weak form of ultraviolet radiation, and although the weakness is beneficial in minimizing the side



8-MOP BINDS TO DNA after being activated by ultraviolet radiation. 8-MOP (8-methoxypsoralen), shown at the top, is a simple compound consisting of two structures: a furan ring and a coumarin. On reaching the cell nucleus, it slides between the paired bases of the DNA chain (*1*). After absorbing a photon of UVA, the molecule forms a pair of bonds with a nucleotide base on one DNA strand (*2*). After absorbing a second photon, 8-MOP can bind to a base on the other strand of DNA (*3*). Such linkage of the strands in the DNA helix prevents the DNA from replicating.



PHOTOPHERESIS entails removing one unit of blood (about the amount given in a single donation of blood). The photopheresis machine, shown at the right, was developed by engineers of Therakos, Inc. (a subsidiary of Johnson & Johnson), in collaboration with the author. After centrifugation to separate the blood into its components, the plasma (liquid portion of the blood) and the leukocytes (white blood cells) are combined with saline. Once in the machine, a thin film of the white-cell suspension passes through a clear plastic channel between twin banks of high-intensity UVA lamps. After the irradiation the erythrocytes (red blood cells) are recombined with the remainder of the blood, which is retransfused into the patient.

effects of exposure, it does imply that the channel in which the blood is exposed must be extremely thin. If it is not, the radiation will be largely absorbed by red blood cells before it reaches the leukocytes at the center of the channel. Indeed, in the third-generation machines for the procedure, which has come to be called photopheresis, a film of blood only one millimeter thick is passed between two high-intensity UVA sources.

In the course of treatment about 500 milliliters of blood that has been separated into its components passes by the light source—about the same volume as is given in a single blood donation. From the dosage work already done we knew the amounts of light and drug a *T* cell had to be exposed to, on the average, to inactivate it. In order to determine the length of time the blood needed to be exposed in the apparatus to achieve those levels, my colleagues Francis P. Gasparro and Regina M. Santella and I formulated monoclonal antibodies to the complex of 8-MOP and DNA. A preparation of monoclonal antibodies binds to one and only one type of molecule, and by employing such a preparation we were able to find how much exposure was required to yield the needed number of 8-MOP–DNA bonds in each *T* cell. Our results showed that 150 minutes of UVA exposure in the photopheresis apparatus would be enough.

Because photopheresis was an experimental and potentially dangerous form of therapy, the initial group of patients had to meet stringent criteria. The disease chosen for treatment was the leukemic variant of CTCL. The basis of the disease is a massive expansion of a single clone of *T* cells. There are millions of different clones, or genetically identical populations, of *T* cells in the immune system. The cells of each clone have a receptor capable of recognizing only a single foreign molecule. When that molecule is present in the course of a disease, the corresponding clone of *T* cells expands as part of the immune response. In a malignancy, however, the diseased clone expands massively, dominating the white-blood-cell population and ultimately killing the patient.

In the leukemic type of CTCL the members of the aberrant clone migrate (by way of the blood and the lymphatic vessels) between the skin, which is infiltrated in such a way that the entire body surface is red and swollen, and the other body tissues. This condition is extremely debilitating and also fatal. When patients suffering from the leukemic phase of the disease are treated with standard therapies, they survive on the average for less than three years, succumbing to opportunistic infections or to destruction of vital organs by the malignant cells. Studies by my group and by others had shown that aggressive leukapheresis (two or three times a week for prolonged periods) could yield temporary improvement. Yet the treatment was onerous and the improvements were brief. By 1982, when we began our trial of photopheresis, no major medical center was any longer performing it with regularity for treating CTCL.

We began the treatments with a conservative regimen aimed only at establishing the toxicity of the new procedure. At intervals of a month, patients underwent photopheresis on two successive days. Because so many of the malignant cells were harbored in the skin, the lymph glands and the spleen, such infrequent treatments of the blood would make it impossible to irradiate any more than 10 percent of them. It was therefore anticipated that once the procedure had been found safe, the frequency would have to be increased greatly. Imagine our surprise when it was discovered that four of the first five patients responded after only from six to 10 treatments and that eventually the severely involved skin of two cleared completely.

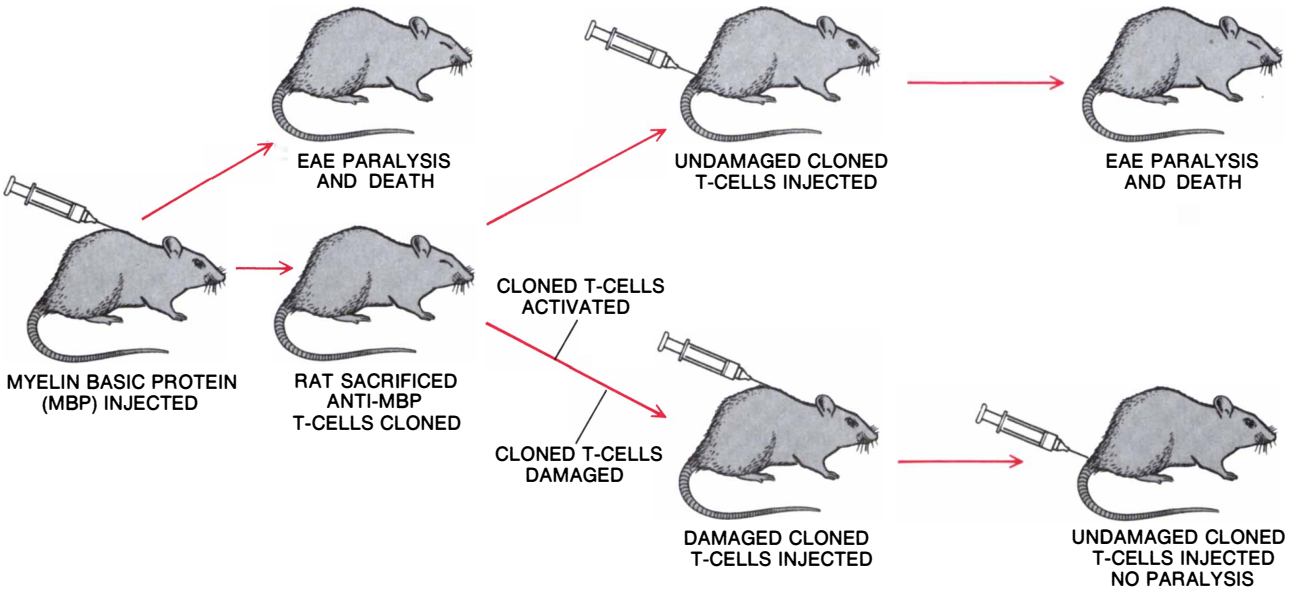
This was a medical mystery that demanded explanation. One thing seemed clear from the outset: the superiority of photopheresis over leukapheresis lay in the fact that after being damaged the white blood cells, rather than being discarded, were returned to the patient's system. This aspect of photopheresis, which had initially generated the most apprehension, seemed to lie at the heart of its efficacy. But why?

No answer was immediately forthcoming, but in the meantime a much larger trial was organized with the advice of the FDA. A total of 37 patients unresponsive to standard therapy were enrolled at the medical centers of Yale University, Columbia University, the University of Pennsylvania, the University of California at San Francisco, the University of Vienna and the University of Düsseldorf. Of that number, 27 responded, including 20 of 26 with lymph-node involvement (indicating advanced disease). Not only is this level of response remarkable in a disease resistant to standard treatments, but also side effects were minimal; photopheresis has now been approved by the FDA as a standard form of therapy for the advanced form of CTCL.

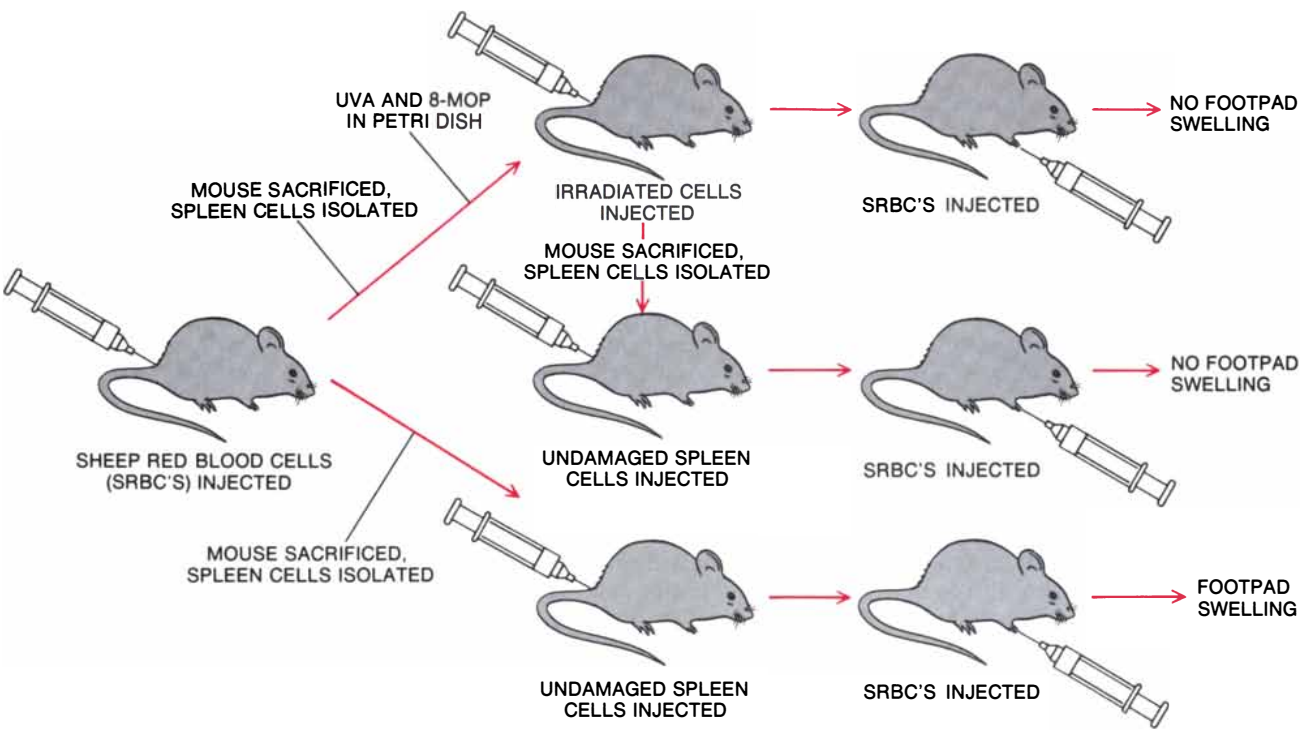
To my collaborators and me in our role as clinicians this was a satisfying accomplishment. Yet as research workers we continued to be nagged by the puzzle of why the treatment was so effective. Perhaps the most significant finding of the expanded multicenter trial was that the patients who responded best were those whose immune systems were strongest at the outset. An analysis by my colleague Carole Berger showed that all the patients who responded well had enduring decreases in the number of malignant *T* cells. The normal *T* cells damaged by photopheresis, on the other hand, were readily replaced by the body's reserves. Somehow the light-and-drug therapy resulted in selective destruction of the malignant clone.

That was a paradoxical result, however, because the treatment itself was clearly nonspecific: all the white blood cells were exposed on the average to the same quantity of 8-MOP and the same amount of radiant energy. The most plausible resolution of the paradox was that there were so many more members of the malignant clone than of any other *T*-cell clone that when they were returned to the body, the damaged diseased cells elicited a specific, intense response whereas the others did not. A malignant clone frequently expands enough so that its members outnumber those of any other single clone by, say, a million to one. Hence when they are returned to the body, the spleen will mount a potent immune response to them but not to the normal cells, which are too infrequent to elicit the same powerful response.

But how, one might ask, can the immune system respond to a group of its own *T* cells, malignant or not? The answer takes us out into the deep waters of modern immunology and specifically toward the work of Irun Cohen [see "The Self, the World and Autoimmunity," by Irun R. Cohen; SCIENTIFIC AMERICAN, April]. The receptor that enables a *T* cell to recognize a specific antigen, or foreign molecule, is itself a protein that can be recognized by the immune system just as the protein of a pathogen is. The distinctive part of a receptor that enables it to fit with its antigen is called the idiotype; the fit between that same receptor and a second receptor on another lymphocyte that in turn fits it is called an anti-idiotype interaction.



"VACCINE" AGAINST AUTOIMMUNITY, discovered by Irun R. Cohen of the Weizmann Institute of Science in Israel, shares a mechanism with photopheresis. Laboratory rats injected with a protein called myelin basic protein develop an autoimmune disease called experimental autoimmune encephalitis, which is a reaction against myelin in their own nervous systems. The reaction is mediated by *T* cells bearing receptors for basic protein. Some rats die from the disorder, but others recover. The *T* cells responsible for the autoimmune response are removed, cloned, activated and damaged by pressure or chemicals. When the *T* cells are injected into new rats, the new rats are prevented from getting the disorder. If the *T* cells are injected without activation and damage, they do cause the disorder. Although it is promising, this method has drawbacks for use in human therapy. In human diseases it is currently quite difficult or even impossible to isolate the relevant clone of *T* cells. What is more, the damage and activation steps may lead to side effects when the cells are restored to the patient.



AUTHOR'S EXPERIMENT shows 8-MOP and UVA can prevent a specific immune response by a mechanism analogous to the one discovered by Cohen. Mice injected with sheep red blood cells (SRBC's) develop an immune response against the sheep cells. The cells responsible for the reaction are *T* cells concentrated in the spleen. The spleen cells can be removed and subjected to UVA and 8-MOP in a laboratory dish. When they are put back into other mice, they specifically prevent the immune response to SRBC's. In contrast to the method shown in the upper illustration, no cloning and no additional activation are required. Hence light and 8-MOP can provide the basis for a method—photopheresis—that is already clinically practical.

Since the immune system as a whole is not damaged by photopheresis, the most attractive possibility to explain our results is that the immune system (in particular the spleen) mounts a powerful response by means of cells bearing anti-idiotypic receptors recognizing the malignant clone.

These observations suggested that our results might best be interpreted in the light of animal studies by Cohen and his group. They worked with several experimental diseases, in which an autoimmune response can be evoked by inoculating a laboratory animal with a specific substance. The basis of the disease is that the injected antigen resembles molecules found in the body's tissues, and as a result the *T* cells against the antigen also attack normal tissue. For example, rats inoculated with basic protein (a component of the myelin that sheathes certain nerves) develop a paralytic disease of the nervous system called experimental autoimmune encephalomyelitis, or EAE.

Cohen's most dramatic finding was that if one could identify the clone of *T* cells responsible for the autoimmune response, that clone could be employed as a "vaccine." The anti-basic-protein clone, for example, was removed from rats that had had EAE and recovered. The cells were subjected to the effects of certain activating substances, and the receptors on their surfaces were aggregated by various means to render them more immunogenic. When the treated cells were reinjected into other rats, those animals were protected against developing EAE: the treated cells acted as a "vaccine" against autoimmunity.

Those results seemed to constitute a remarkable parallel to our own. We had, after all, removed *T* cells from the body, damaged them and then reinserted them. Moreover, many of the cells we worked with came from one particular clone—the malignant one. Perhaps photopheresis was "vaccinating" the CTCL patients against their own cancer.

To find out whether that was the case, several experimental approaches mimicking the photopheresis procedure were tried in our laboratory by Berger, Maritza Perez and Liliane Laroche. In one such approach an inbred strain of genetically identical mice were inoculated with red blood cells taken from sheep, which produce a strong immune response. Spleen cells from these mice were removed and exposed to 8-MOP and ultraviolet radiation in laboratory dishes. The spleen cells were then injected into mice of the same inbred strain. When the mice so "vaccinated" were exposed to sheep red blood cells, they showed no immune response. The likeliest reason is that the "vaccination" caused the mice to mount a strong anti-idiotypic immune response against the very cells from the spleen that would ordinarily defend against the foreign cells.

That result (and several other analogous ones) made it clear that photopheresis could lead to the effect observed by Cohen and his colleagues. Indeed, with the clairvoyance of hindsight we now see that 8-MOP and UVA are ideal means for achieving such effects. The capacity for activating a potent drug so that it injures large numbers of abnormal *T* cells bearing a specific surface marker before they are returned to an intact immune system is most propitious for stimulating a strong immune response specifically against those cells. Now it is possible to understand why photopheresis is so effective: the damaged cells of the malignant clone in effect prime the immune system to destroy that clone specifically, ridding the body of its cancer.

An important theoretical question that remains to be resolved is why, if there are so many cells of the malignant clone in the patient's blood to begin with, the immune system does not mount the same healing response on its own. After all, in leukemia the malignant clone often accounts for more than 20 percent of all leukocytes. Cohen's work suggests that in the animal systems the *T* cells must be first activated and then damaged before they will elicit the desired response. The malignant *T* cells are already activated (as part of the cancer process), but it seems that the photopheresis is necessary because the damage inflicted on the cell somehow increases its immunogenicity. Work is now proceeding in our laboratory and in others to find out just why that is so.

Photopheresis may ultimately be applied to treating a wide variety of diseases that entail expansion of specific *T*-cell clones. By the time it is so applied, however, photopheresis may have been improved by a number of modifications. One is the improvement of the drug itself. A synthetic analogue of 8-MOP known as aminomethyl-trimethylpsoralen is more soluble in water than the natural substance and also has a greater affinity for DNA. Those properties lead, as Gasparro and his co-workers have shown, to significantly greater activity per molecule. In the future other drugs well suited to photopheresis will no doubt be found or synthesized.

In addition, there will certainly be better and more specific ways of delivering the drug and the accompanying light to target tissues. One experimental delivery system has been the subject of work in our laboratory by Shrishailam Yemul, Alison Estabrook, Hagen Bayley, Berger and me. In this system monoclonal antibodies are attached to the surface of vesicles (small sacs) consisting of the same type of lipid bilayer as the cell's own outer membrane. By choosing the appropriate antibody, we have been able to cause the vesicles to bind preferentially to particular types of cells. Molecules of the photoactive drug pyrene are incorporated into the vesicles; when the vesicle reaches its target and the pyrene is activated by light, the drug blows holes in the cell's outer membrane, killing the cell.

Several other experimental systems are also being worked on, all offering theoretical advantages of very high specificity. In these systems the drug could be activated not only outside the body, as in current photopheresis systems, but also within the body if a way can be found to deliver light to where it is needed. The hybrid molecules in such systems can be considered prototypes of drugs belonging to a new category of pharmacological agents: those activated by light. Current technical knowledge makes it possible to modify almost any drug in such a way that it can be activated only by light. Hence we may one day look back on the unmodified psoralens as little more than harbingers of an entirely new class of therapeutic agents.

FURTHER READING

CUTANEOUS T CELL LYMPHOMA. Robert M. Knobler and Richard L. Edelson in *The Medical Clinics of North America,* Vol. 70, No. 1, pages 109–138; January, 1986.

SELECTIVE KILLING OF T LYMPHOCYTES BY PHOTOTOXIC LIPOSOMES. Shrishailam Yemul, Carole Berger, Alison Estabrook, Sylvia Suarez, Richard Edelson and Hagen Bayley in *Proceedings of the National Academy of Sciences of the United States of America,* Vol. 84, No. 1, pages 246–250; January, 1987.

TREATMENT OF CUTANEOUS T-CELL LYMPHOMA BY EXTRACORPOREAL PHOTOCHEMOTHERAPY: PRELIMINARY RESULTS. Richard Edelson et al. in *The New England Journal of Medicine,* Vol. 316, No. 6, pages 297–303; February 5, 1987.

Perceiving Shape from Shading

Shading produces a compelling perception of three-dimensional shape. One way the brain simplifies the task of interpreting shading is by assuming a single light source

Vilayanur S. Ramachandran

Our visual experience of the world is based on two-dimensional images: flat patterns of varying light intensity and color falling on a single plane of cells in the retina. Yet we come to perceive solidity and depth. We can do this because a number of cues about depth are available in the retinal image: shading, perspective, occlusion of one object by another and stereoscopic disparity. In some mysterious way the brain is able to exploit these cues to recover the three-dimensional shapes of objects.

Of the many mechanisms employed by the visual system to recover the third dimension, the ability to exploit shading is probably the most primitive. One reason for believing this is that in the natural world many animals have evolved pale undersides, presumably to make themselves less visible to predators. "Countershading" compensates for the shading effects caused by the sun shining from above and has at least two benefits: it reduces the contrast with the background and it "flattens" the animal's perceived shape. The prevalence of countershading in a variety of species, including many fishes, suggests that shading may be a crucial source of information about three-dimensional shape.

Painters, of course, have long exploited lighting and shading to convey vivid illusions of depth. Psychologists, however, have not devoted much research to uncovering the mechanisms by which the eye and the brain actually take advantage of shading information. My colleagues and I therefore embarked on a set of experiments intended to reveal what some of the mechanisms might be.

We started out by creating a set of computer-generated displays of simple objects in which subtle variations in shading alone convey the impression of depth. We made sure the images were devoid of any complex objects and patterns, because our goal was to isolate the brain mechanisms that process shading information from higher-level mechanisms that may also contribute to depth perception in real-life visual processing.

Our experiments were based on circular, shaded shapes that create a compelling sensation of depth [*see* a *in upper illustration on page 78*]. The shapes either pop outward like eggs or inward like the cavities of an egg carton. The shapes are ambiguous because the brain does not know from which direction the light is shining. With some effort you can mentally shift the light source to invert the depth of the objects.

Intriguingly, when you mentally reverse the depth of one object, all the other objects in the display reverse simultaneously. This raises an interesting question: Is the propensity for seeing all objects in the display as being simultaneously convex (or concave) based on a tendency to see all of them as having the same depth or is it based on the tacit assumption that there is only one light source? To find out, we created a display in which objects in one row are mirror images of objects in the other row [*see* b]. In this display, when subjects see one row of objects as convex, they always perceive the other as concave.

We drew two conclusions from this simple experiment. First, the derivation of shape from shading cannot be a strictly local operation; it must be a global process involving either the entire visual field or a large portion of it. Second, the visual system seems indeed to assume that only one light source illuminates the entire image. This may be because our brains evolved in a solar system that has only one sun.

Another manifestation of this rule is seen in a complex shape suggesting a white tube lit from the side [*see* c]. The shape nearly always appears convex, perhaps because of subtle cues such as the occlusion of one part of the tube by another, or because of a general tendency to see such shapes as convex. Interestingly, the depth of the two disks superposed on the tube is no longer ambiguous; one is clearly a bump and the other a cavity. Apparently certain features of an object can inform the brain about the direction of illumination, and the depth of other parts of the object are then made to conform to the light source.

The visual system not only assumes a single light source but also tends to assume, naturally enough, that the light comes from above. We vividly demonstrate this effect with a display in which one group of shaded circles is simply the upside-down version of another group [*see lower illustration on page 78*]. Subjects always perceive group *a* as consisting of spheres and group *b* as consisting of cavities. If you turn the page upside down, you will find a striking reversal of depth: the objects in group *b* now appear convex and those in group *a* appear concave. (You can amuse yourself by cutting out the illustration and mounting it on a turntable. How fast can you spin the turntable before you stop seeing the reversals?)

These observations suggest that the brain assumes the sun shines from above. But how does the brain know

VILAYANUR S. RAMACHANDRAN is professor of psychology at the University of California, San Diego, and has a joint appointment as a visiting associate in biology at the California Institute of Technology. He obtained an M.D. at the University of Madras and a Ph.D. in neurophysiology from the University of Cambridge. He has held visiting appointments at several institutions, including the University of Oxford, where he studied the development of binocular neural mechanisms in cats and sheep. His current research goal is to find physiological mechanisms underlying the perceptual effects described in the article.

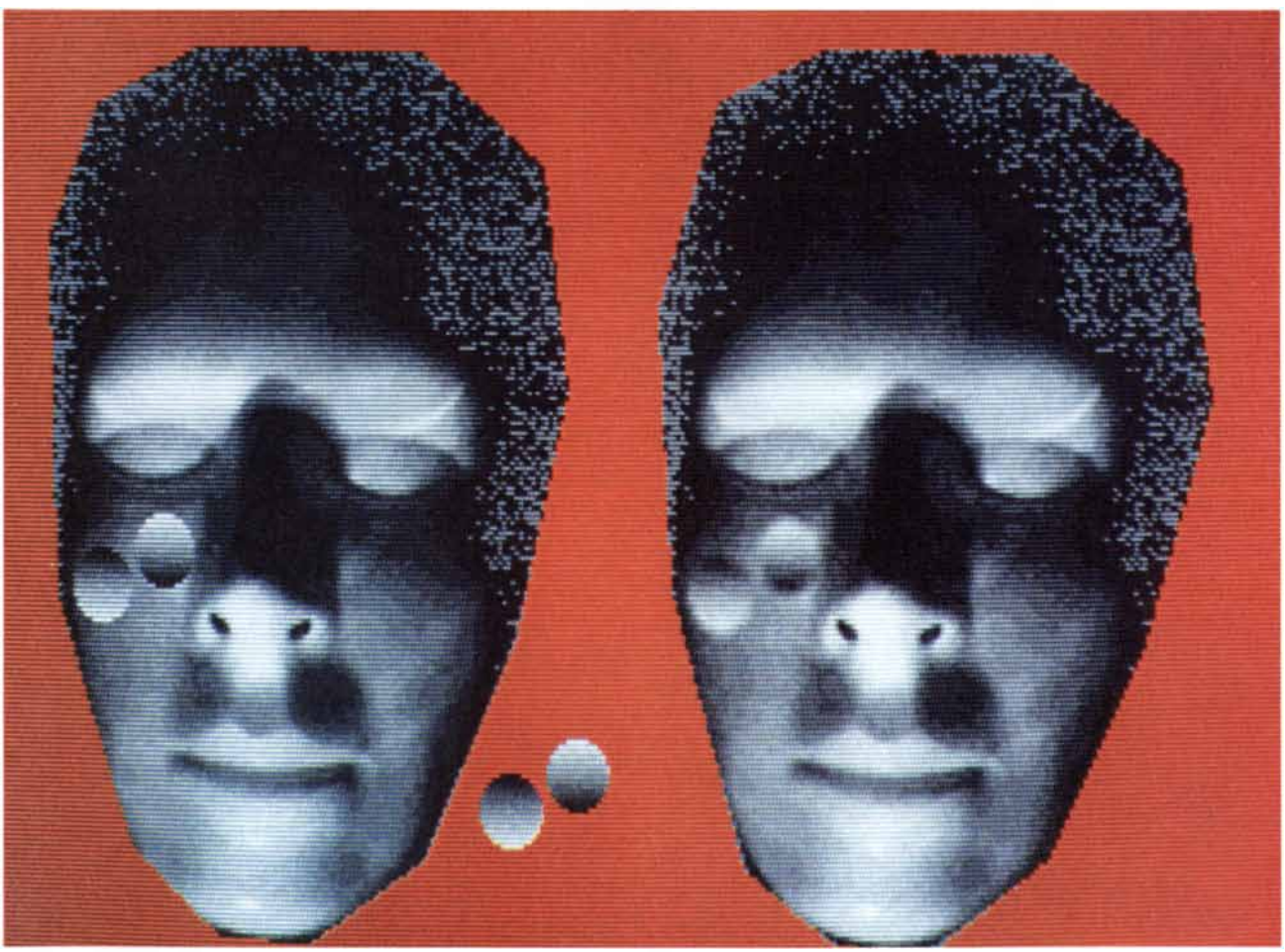
"above" from "below"? Is it the object's orientation in relation to the retina that matters, or is it its orientation with respect to the external world? To appreciate this point try the following experiment. Lie on a couch and let your head hang over the edge so that you are looking at the world upside down. Now ask a friend to stand behind your head and hold the lower illustration on the next page upright. The objects in group *a* will look concave and those in group *b* convex; that is, you get the same effect as you did when you rotated the page. Thus it is the orientation of the object on the retina that matters. Your objective knowledge of up and down does not affect your perception of depth.

Shading by itself generates only a weak impression of three-dimensional shape. To convey a convincing impression of depth the shaded surface must also be enclosed by an outline. Indeed, in many of our displays the luminance variation only roughly approximates the smooth, cosine variation of true shading, and yet the mere presence of a circular outline around the shaded region can generate a compelling illusion of a spherical surface. This raises a new question: What is the exact role of the outline in determining the perception of shape from shading?

To answer the question, we designed a pair of objects that have the same shading but different outlines [*see illustration on page 79*]. Both images have the same luminance gradient: a photocell dragged across each image would register identical variations in the distribution of luminance. Yet the images are strikingly different. The upper image suggests three cylinders lying side by side, whereas the lower image conveys the unmistakable impression of a sheet of corrugated metal. The perceptions seem to depend completely on the contours of the top and bottom edges of the surfaces.

We conclude from these demonstrations that when shading cues are ambiguous, information from borders helps to resolve ambiguity throughout the image. Interestingly, the perceived location of the light source also shifts to conform to the perceived surface. In the upper image on page 79 the light seems to originate perpendicular to the page whereas in the lower image the illumination is from the far left or the far right. It is remarkable that changing an object's boundaries can produce such striking changes in perception.

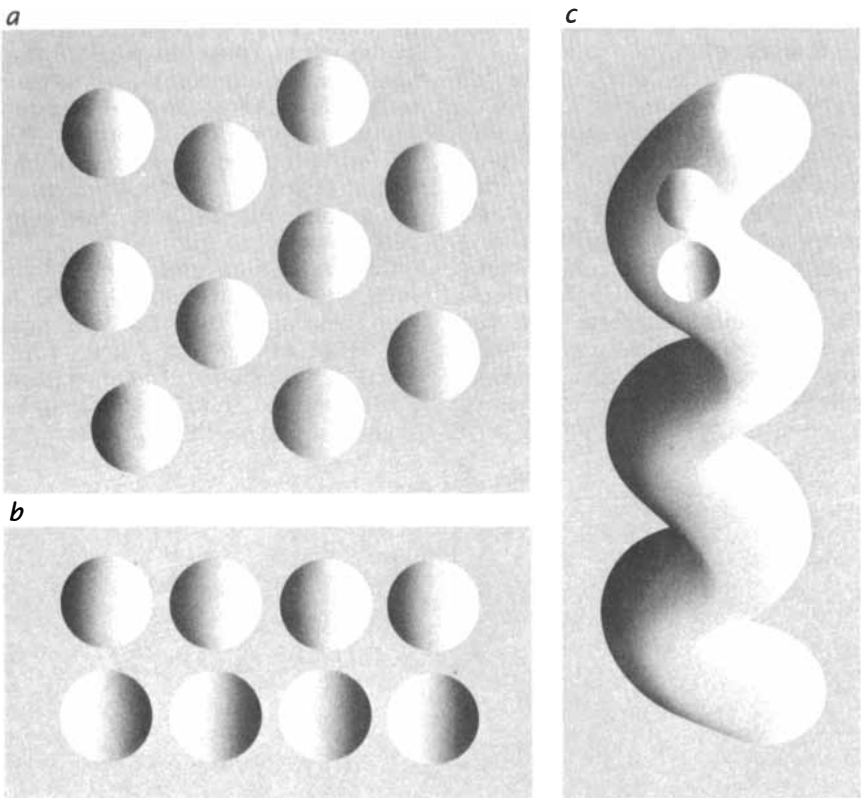
Our next demonstration shows that even illusory contours will work. A typical example consists of four dark gray disks with a "bite" taken out of each one [*see top illustration on page 80*]. When the disks are in proper alignment, one has the impression of a



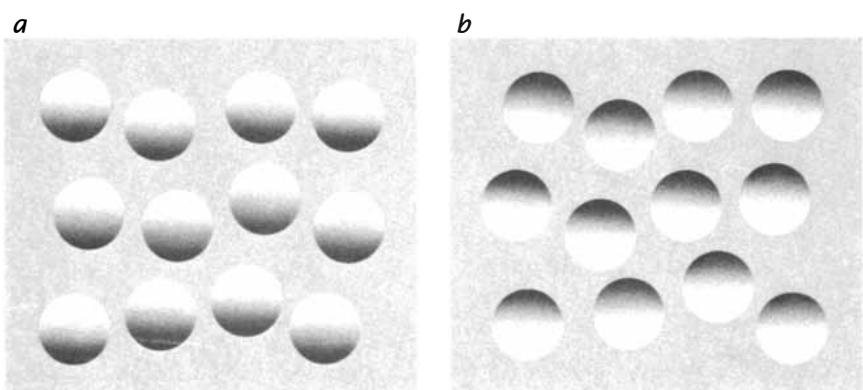
HOLLOW-MASK INTERIORS lit from above produce an eerie impression of protruding faces lit from below. In interpreting shaded images the brain usually assumes light shining from above but here it rejects the assumption in order to interpret the images as normal, convex objects. Notice the two disks near the chin still appear as though lit from above: the right disk seems convex and the left one concave. When the disks are pasted on the cheek (*left*), their depth becomes ambiguous. When blended into the cheek (*right*), the disks are seen as being illuminated from below, like the rest of the face.

large pale disk at the center partially occluding the gray disks. Indeed, faint lines seem to connect the concave edges of the gray disks, although such lines do not exist physically.

What happens if we replace the background of this display with one in which the luminance varies from top to bottom? The new display looks flat at first, but on prolonged viewing the region inside the illusory disk starts to bulge out toward the observer and may even detach itself from the background to take on the appearance of a floating sphere. Oddly enough, an illusory contour seems to work even better than a real outline. The reason is not entirely clear but the result suggests that the brain regards partial occlusion as stronger evidence for the existence of an object than the presence of a mere outline. After all, the outline might equally well depict a loop of thin wire or a transparent soap bubble.



SPHERES OR CAVITIES? It depends on where you think the light source is. You can reverse the depth of the objects (*a*) by mentally shifting the light source from left to right. In a second array (*b*) each row by itself is ambiguous, but once you see one row as convex the other row will always appear concave. It is almost impossible to see both as simultaneously convex or concave. The convoluted form (*c*) suggests a white tube lit from the right. The two disks on it seem to conform to the lighting scheme; the top disk is seen as a bump and the bottom one as a cavity. The last experiment was done by the author in collaboration with Dorothy Kleffner and Steven J. Cobb.



BRAIN ASSUMES light comes from above. The objects in group *a* therefore appear convex, whereas those in group *b* appear concave. If you turn the page upside down, the objects will reverse in depth. By turning your head upside down and looking at the page, you can prove it is the orientation of the pattern on the retina that matters.

Another way the visual system delineates objects is by changes in surface reflectance, or the proportion of light reflected by surfaces. A photocell moving across an object's border will usually register an abrupt shift in luminance. What would happen if the outline were defined by a change of color rather than a change of luminance? We took a typical shaded "sphere" and replaced the homogeneous gray background with a colored background in which the luminance gradient matched that of the sphere. The result was dramatic: the illusion of depth dissolved and the sphere appeared flattened, even though its outline was distinctly visible because of the contrast in hue. We concluded that the shape-from-shading system cannot make use of edges defined by color differences. One reason may be that our primitive primate ancestors, which resembled tarsiers, were nocturnal and color-blind; in their twilight world they relied on luminance contrast alone to perceive depth.

These demonstrations imply that the brain recovers information about the shape of objects by combining outlines and shading cues. What does the brain do with the shapes once it has recovered them? An important capacity of perception is the ability to segregate figure from ground. Even in a cluttered scene the visual system can easily decide which features in the image belong together to form objects. In a high-contrast photograph one can see a Dalmatian

dog against a dappled background [*see bottom illustration on next page*]. Similarly, one can mentally "lift out" a group of lines that have a particular orientation from a field of lines with a different orientation. On the other hand, it is impossible to segregate a group of mirror-reversed letters from unreversed ones.

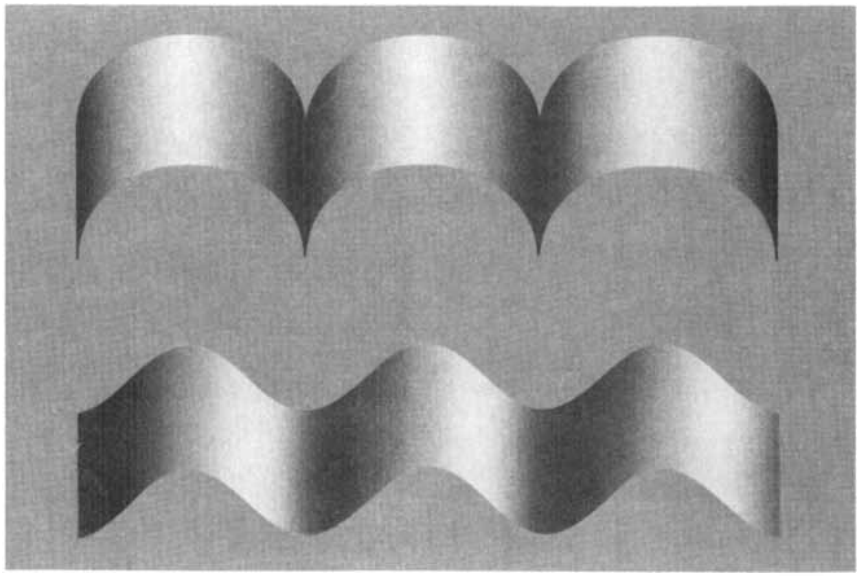
The laws of perceptual grouping were first studied systematically by Anne M. Treisman of the University of California at Berkeley, Bela Julesz of the AT&T Bell Laboratories and Jacob Beck of the University of Oregon. These investigators discovered several important principles. First, they found that an important early stage of visual perception involves extracting certain elementary features, which Julesz calls textons. Examples include oriented edges, color and direction of movement. Once the visual system has extracted the elementary features, similar features are grouped together to form objects. Indeed, Beck suggests that only elementary features, by definition, can be grouped in this way. Presumably, then, alphabetic characters are not elementary features as far as the visual system is concerned.

What about three-dimensional objects, though? Our next several demonstrations show that even shapes defined exclusively by shading can serve as elementary features of visual perception. In an array of cavities interspersed with convex shapes, for example, the convex shapes can mentally be grouped together to form a separate depth plane that is clearly segregated from the concave shapes in the background [*see illustration on page 81*].

When one views this display, it appears as though the visual system passes through several stages of processing. In the earliest stage the system performs computations for defining the three-dimensional shapes, taking several seconds. Once the convex shapes have emerged, one has the distinct impression of being able to "hold on" to them indefinitely in order to group them with similar items in the display. Finally, after the objects are grouped, they are clearly segregated from irrelevant items in the background. The extraction and grouping of textons, then, although usually described as a one-step operation, may in fact involve several distinct perceptual capacities that act together to delineate figure from ground.

We wondered whether the perceptual grouping observed in that display might be the result of some other, more elementary feature than the three-dimensional shape. For example, because the convex shapes differ from the concave ones in the polarity of their bright-to-dark luminance, one might suppose the grouping is achieved by latching on to luminance polarity. To rule out this possibility, we created a display of objects that have the same luminance polarities as those in the preceding display but that do not carry any depth information. It is virtually impossible to achieve perceptual grouping in this display. Even after you have spotted all the targets individually, you will not be able to segregate them from the rest of the objects. Clearly the grouping observed in the preceding display must be based on three-dimensional shape rather than on luminance polarity.

I have pointed out that the illusion of depth is much more powerful when the illumination is from above than when it seems to come from the side. Similarly, lighting from above greatly enhances one's ability to group and segregate images. You can verify this by simply rotating group *a* on page 81 by 90 degrees: the impression of depth will diminish and there will be a considerable reduction in perceptual segregation. This further supports the idea that perceptual grouping must be based on three-dimensional shape. Moreover, these groupings can themselves represent higher-level shapes, such as a triangle. It might be interesting to employ stimuli of this kind to find out whether infants and brain-damaged patients can perceive shape from shading; for example, would an infant respond to spheres arranged to suggest a face?

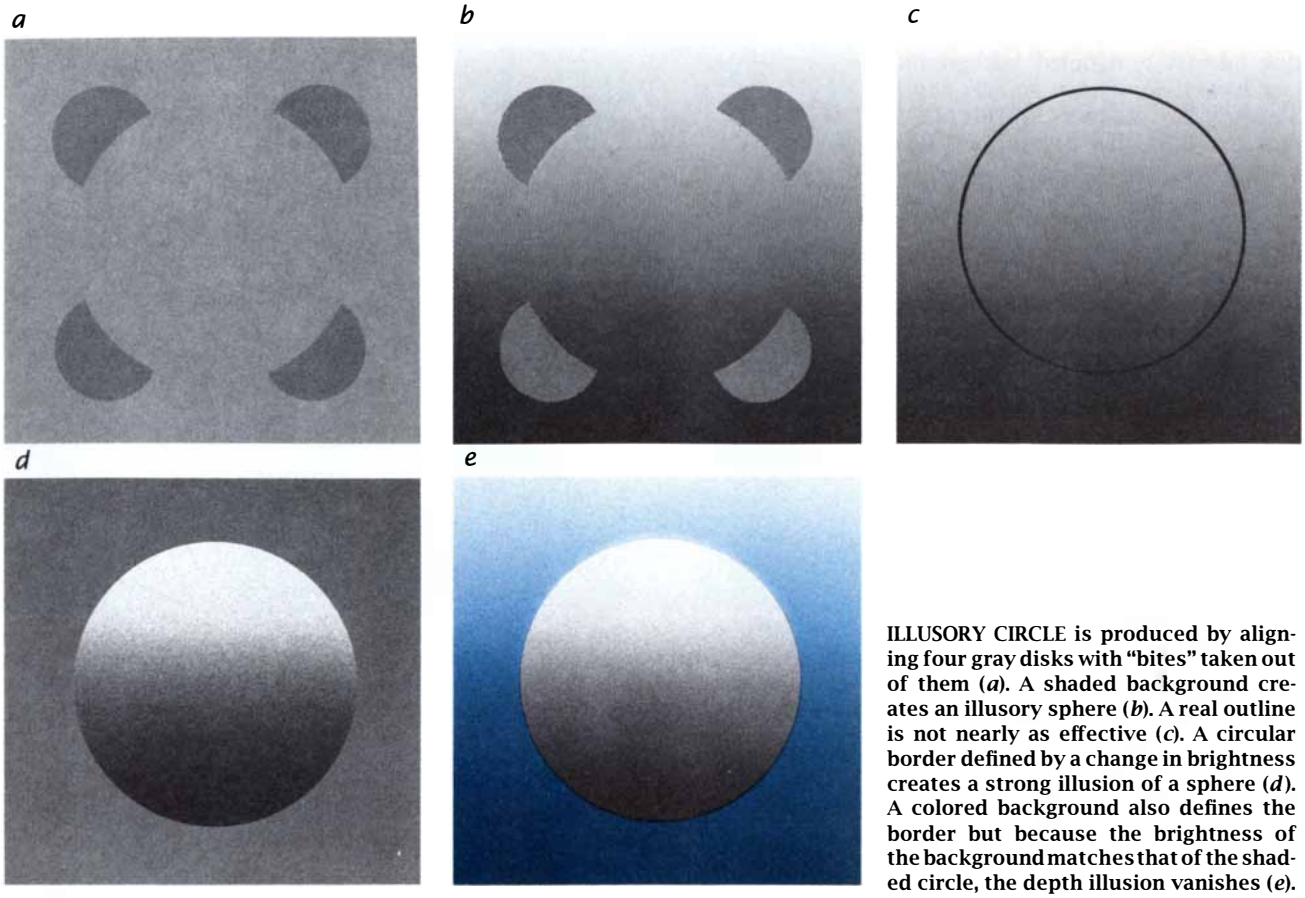


BOUNDARIES influence the interpretation of shaded surfaces. Both images have the same shading variation but the top image suggests three cylinders lit vertically to the page and the bottom one a corrugated metal sheet lit from far left (or far right).

Another remarkable capacity of visual perception is the ability to detect symmetry. This ability extends to fairly complicated shapes, such as plants, faces and Rorschach inkblots. How does the visual system detect symmetry? Does it match all the individual features on one side with those on the other side to determine whether an object is symmetrical? Or does it group features into more meaningful shapes and then look for symmetry in those shapes? Our next demonstration is an attempt to answer these questions.

We compared two arrays of shaded circles [*see top illustration on page 82*]. Subjects usually perceived the left-hand array as spheres and cavities arranged symmetrically about a horizontal axis. Yet a point-by-point examination reveals that the bottom half of the array is not a mirror image of the top half. In fact, it is the array on the right that is truly symmetrical. These results imply that the perception of symmetry is based on three-dimensional shape rather than on the simple distribution of bright and dark areas in the image. You can verify this by rotating the illustration 90 degrees to eliminate the strong impression of depth. You will now see that the right-hand array is more symmetrical than the left-hand one.

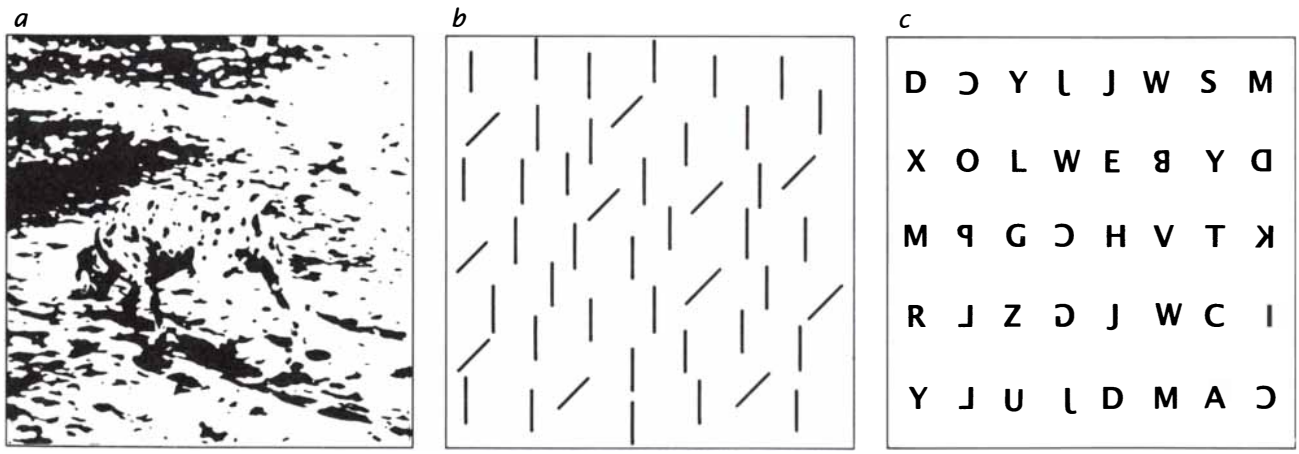
Our observations suggest that shading information is extracted fairly ear-



ILLUSORY CIRCLE is produced by aligning four gray disks with "bites" taken out of them (*a*). A shaded background creates an illusory sphere (*b*). A real outline is not nearly as effective (*c*). A circular border defined by a change in brightness creates a strong illusion of a sphere (*d*). A colored background also defines the border but because the brightness of the background matches that of the shaded circle, the depth illusion vanishes (*e*).

ly in visual processing. Indeed, there may even be neural channels specifically committed to the purpose. Recently Terrence J. Sejnowski and Sidney R. Lekhy of Johns Hopkins University raised the possibility that such cells may exist, based on work with a computer simulation. They began with a "neural network" consisting of three layers of cells: an input layer, a hidden layer and an output layer. Input-layer cells were modeled on the circular, "center surround" receptive fields of cells in a cat's eye. A learning algorithm adjusted the strength of signals passing from cells in one layer to the next, and after 40,000 trials the network could correctly associate shaded shapes with their three-dimensional axes of curvature.

What happened next came as a surprise: the investigators examined the responses of the cells in the hidden layer and found that they responded to bars of various lengths, widths and orientations, bearing an uncanny resemblance to edge-detector cells found in the visual cortex of cats and monkeys. Intriguing as this computer simulation is, its biological relevance is still not clear because the investigators deliberately excluded outlines and other cues known to play a crucial role in human vision. It remains to be seen whether the resemblance between the hidden units and the cor-



PERCEPTUAL GROUPING of elementary features enables one to segregate the shape of a Dalmatian dog from a speckled ground in a high-contrast photograph (*a*). Slanted lines can be grouped and envisioned as occupying a separate plane from vertical lines (*b*). Mirror-reversed letters, however, cannot be visually grouped and segregated from normal letters (*c*).

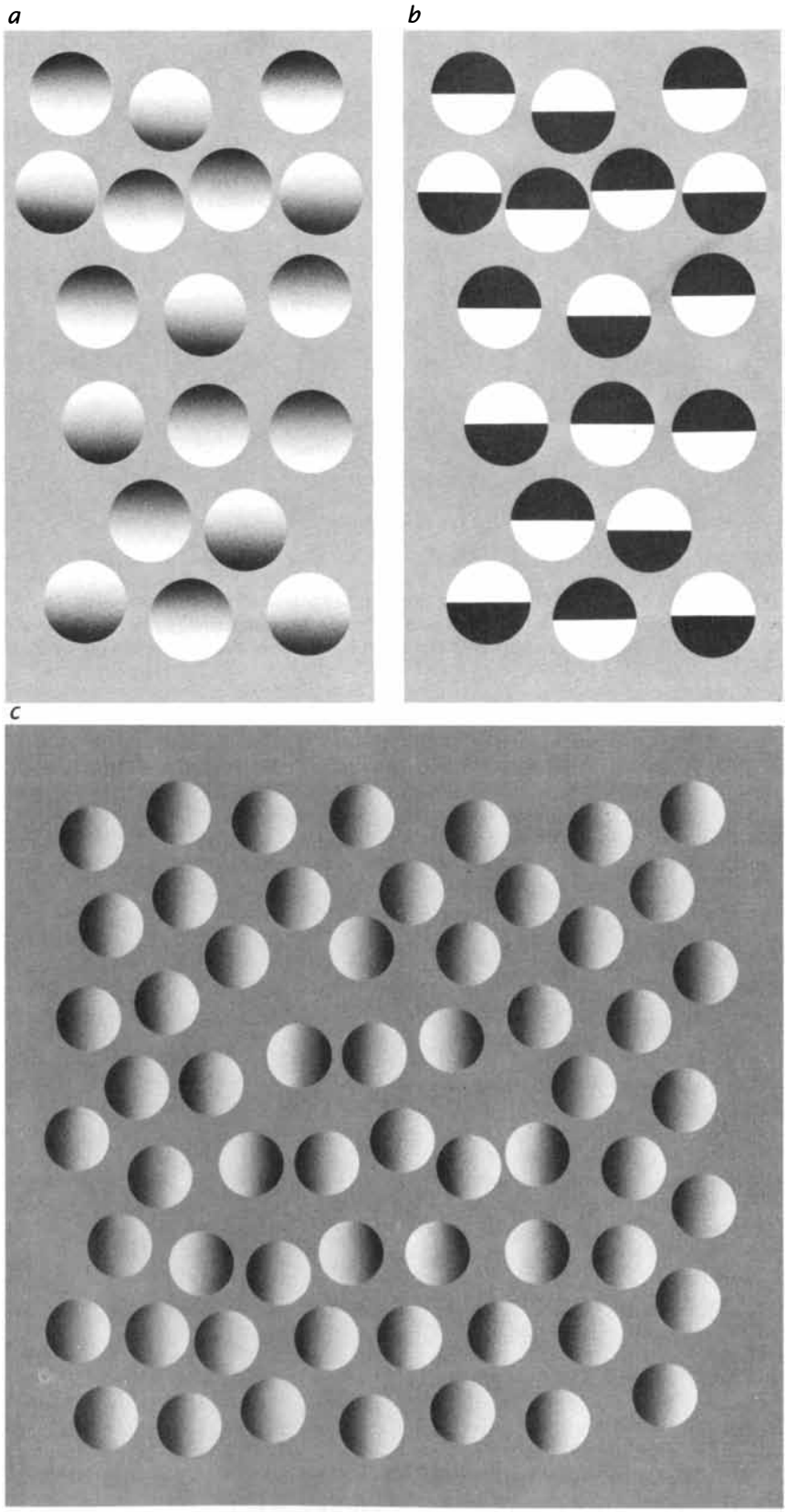
tical edge detectors is merely a coincidence or whether edge-detector cells actually serve to extract three-dimensional shapes from shading.

I have so far considered stationary images, but what about moving objects? In nature it is a reasonably safe bet that anything that moves is either prey or predator. Consequently the visual system appears to have evolved a wide variety of mechanisms for detecting movement. Evidence suggests that the ability to see movement is mediated by specialized groups of brain cells. Can the brain mechanism enabling us to perceive motion also take advantage of information provided by shading? In order to find out we decided to exploit a well-known illusion called apparent motion [see "The Perception of Apparent Motion," by Vilayanur S. Ramachandran and Stuart M. Anstis; SCIENTIFIC AMERICAN, June, 1986].

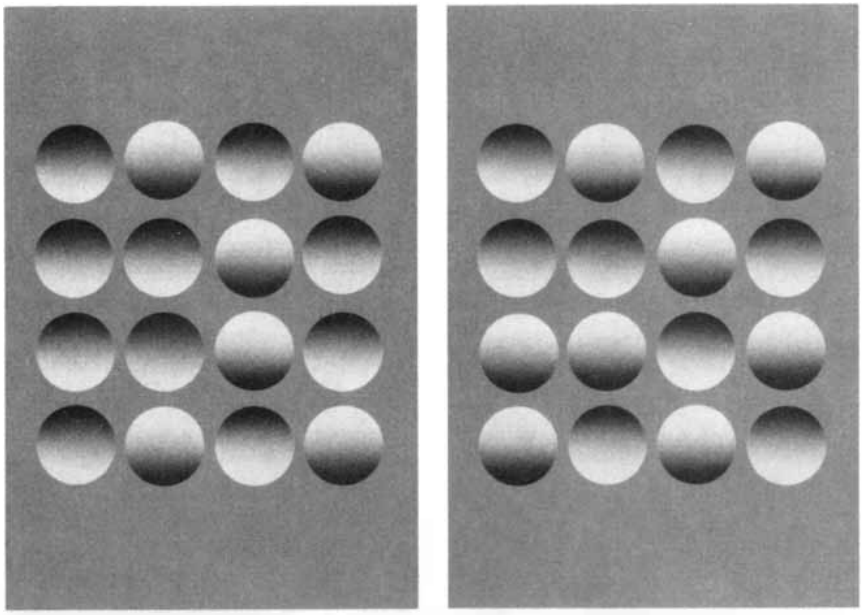
A simple example of apparent motion is produced by flashing two spatially separated spots of light in rapid alternation. Instead of seeing two lights flashing on and off one usually sees a single light jumping back and forth. To investigate the role of shading cues in human motion perception we created a display that alternated rapidly between one frame showing a shaded convex object above a concave one and a second frame in which the objects are reversed. Eleven naive subjects reported seeing a sphere jumping up and down between two holes in the background.

The result suggests that the brain must first compute three-dimensional shape before it can perceive apparent motion. Indeed, subjects often take tens of seconds to develop a depth impression, during which time they see no apparent motion. It therefore seems unlikely that the apparent motion could be based on some other, more primitive feature of the image. To demonstrate the point more directly we rotated the entire display by 90 degrees. This reduced the impression of depth considerably and led to an almost complete loss of the apparent-motion effect.

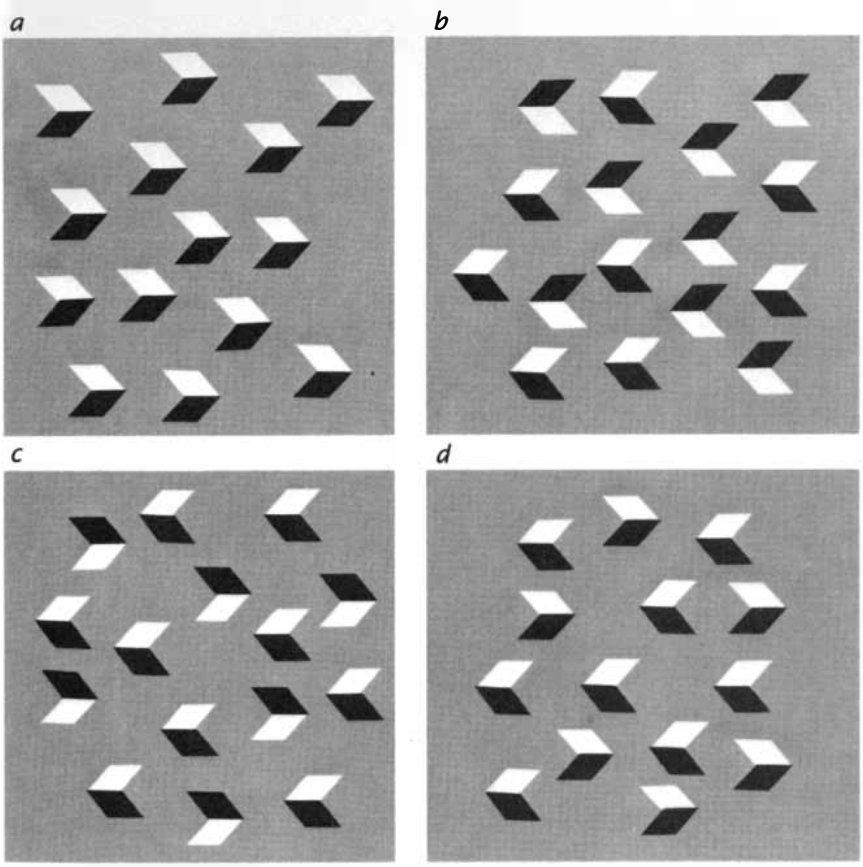
The visual system, then, appears to extract a three-dimensional object from shading cues and to perceive movement based on the three-dimensional image, rather than using the "primitive" two-dimensional image directly. Certain cells in the visual cortex of the monkey respond to the apparent motion of simple stimuli such as the flashing spots of light described above. It might be interesting to see



VISUAL SYSTEM can pick out convex shapes from concave ones and group them together (*a*). In an array that conveys the same luminance polarities as the preceding one but no depth information (*b*) it is impossible to visually segregate objects. In an array lit from the side (*c*), perceptual grouping becomes easier when you rotate the picture by 90 degrees. The convex objects stand out and form a triangle. Shading can define such complex shapes for further visual processing. A similar idea has been proposed by Alex Pentland of the Massachusetts Institute of Technology.



SYMMETRY PERCEPTION occurs after the brain extracts shape from shading. In the left-hand array spheres and cavities seem arranged symmetrically about a horizontal axis. But in two dimensions it is the right-hand array that is truly symmetrical.



CHEVRONS in group *a* appear as illusory cubes or as "gravestones" casting shadows, all lit from the same angle. In group *b* a "pointing" rule seems to override the single-light-source rule and the image is seen as a set of cubes whose faces have different reflectances. Group *c* is always seen as a mixture of cubes and gravestones because this interpretation satisfies the single-light-source rule. Group *d* is ambiguous, and it is difficult to unify the figures into a coherent interpretation.

whether these cells would respond to motion based on objects whose shape is perceived from shading.

Clearly, visual perception relies on a constellation of biological processes to arrive at a three-dimensional representation of the world. In order to create this representation the visual system appears to make a variety of simplifying assumptions, such as the rule that there is only one light source. What happens when the visual system tries to construct a coherent scene out of many disparate fragments? Patrick Cavanagh, Diane Rogers-Ramachandran and I recently did a study to try to answer the question.

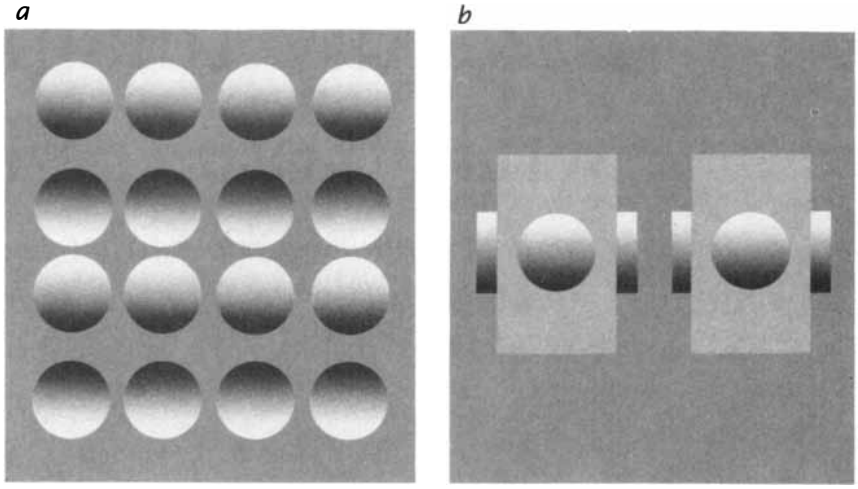
We created simple arrays of randomly placed chevrons, each of which can be viewed as two adjoining faces of a cube [*see bottom illustration at left*]. Array *a* can be perceived as parallel cubes all pointing in the same direction and illuminated by a single light source; the black parallelograms are seen as the shadowed face of the cubes. But equally often the array is perceived as a set of white "gravestones" casting black shadows. By mentally shifting the direction of the light source you can switch from seeing cubes to seeing gravestones. Note that when you see any one figure in the array as a cube you see all others as cubes too. It is impossible, in fact, to simultaneously perceive some figures as cubes and the others as gravestones, because such a perception would violate the single-light-source rule. Interestingly, when the shapes are perceived as cubes, there is a tendency to fill in the missing faces—that is, to perceive illusory surfaces. The illusory surfaces vanish when the shapes are seen as gravestones.

Next we randomly inverted or reversed roughly half of the chevrons in various combinations. These new displays illustrate the subtle interplay of constraints and organizing rules that occurs when the brain tries to create meaningful shapes from isolated fragments. In array *b*, for instance, all the targets usually appear as parallel cubes even though this would be incompatible with a single light source. Apparently when the single-light-source rule cannot be satisfied, it is replaced by a "pointing" rule (or by a rule stating that shapes with similar orientations are in fact parallel surfaces). To avoid conflict the brain simply assumes that the cubes have faces of differing color.

In array *c* you will see a mixture of gravestones and cubes because this

allows the system to satisfy the single-light-source rule. It is actually impossible to see the display as consisting entirely of cubes or of gravestones, because such a perception would not be compatible with either the pointing rule or the light-source rule. You will also find yourself unifying all the items in the array into a single coherent surface so that it suggests a sculptured metal surface with randomly placed "steps" carved out of it. Whereas an ant crawling on the picture would see only chaotic fluctuations in brightness, the human eye surveys the entire image and knits parallel surfaces together to create spatial order and unity.

In array *d* the figures are neither parallel nor able to satisfy the single-light-source constraint. Hence there is a tendency to see the display as a random collection of flat chevrons pointing in opposite directions. Even though an individual figure in the display can sometimes be seen as a cube or a gravestone, it is difficult to unify all of them into a coherent three-dimensional interpretation.



VISUAL IMAGERY, or higher-level information about objects, profoundly influences the perception of shape from shading. Rows of spheres and cavities (*a*) can also be seen as two blurred stripes visible through 16 holes cut in an opaque sheet. One can no longer see the spherical shapes. Each of two pictures depicts a sphere (*b*), but when they are "fused" in a stereoscopic viewer, the spheres vanish and one sees a circular window cut in a rectangular sheet floating in front of a shaded plane.

In the real world, visual imagery—that is, high-level knowledge about what one is seeing—profoundly affects the perception of shape from shading. Indeed, the interaction between visual imagery and perception is one of the most elusive and enigmatic topics in psychology. To illustrate this point, we created an array of shaded circles that on casual inspection appear to form alternating rows of spheres and cavities [*see illustration on this page*]. The display is susceptible to a radically different interpretation, however: it can be seen as a gray sheet with 16 holes cut out, behind which are two blurred, dark stripes. This perceptual switch causes the shaded circles to lose their spherical shape completely.

The tendency to see stripes rather than spheres and cavities can be enhanced by stereoscopic cues. You can see spheres in *b* on this page, but if you were to "fuse" them binocularly through a stereoscopic viewer, you would see a frame with a circular window standing out clearly from a shaded background. Indeed, it becomes virtually impossible to see the circular shape as a sphere instead of a hole. This implies that the extraction of shape from shading is strongly affected by stereoscopic processing.

The interpretation of shape from shading also interacts strongly with the visual system's knowledge of objects, as is strikingly demonstrated by the opening illustration for this article. In these photographs the hollow insides of face masks are illuminated from above; one would therefore expect them to look hollowed out. But the visual system strongly rejects the possibility of hollow shapes and interprets the images as normal faces lit from below. Thus the visual system overrides the assumption of lighting from above in order to be able to interpret the shapes as normal faces.

Now notice the two small, shaded disks between the chins of the two faces. Even though the light on the faces is assumed to come from below, the disk on the right generally is seen as convex and the one on the left as concave—as though they were both illuminated from above. Perhaps the brain treats these objects as being quite distinct from the faces and therefore, in interpreting their shading, adheres to the more "primitive" rule that they are illuminated from above. When the disks are pasted onto the cheek of one of the faces, however, the depth becomes ambiguous: the right-hand disk can appear concave and the left-hand one convex. Finally, when the outlines of the disks are blended into the cheek, they are always seen as being illuminated from below, like the rest of the face. Consequently the disk at the right suggests a dimple and the one at the left looks like a bump or a tumor.

Our research has revealed a variety of rules that are applied early in the visual processing of shape from shading. We have shown that it is possible to trace the flow of information from the very early stages of shape perception to the final stage, where the information interacts with high-level knowledge of light sources and of the nature of complex, three-dimensional objects. The neurological events mediating the process in human beings are still mysterious, but insights from psychology can help to elucidate what these events may be and how they are organized in the brain. New computational models can also offer plausible mechanisms and help to narrow the search. These developments are launching research on visual perception into a new domain, where it may someday be possible to discover the cellular mechanisms in the brain that enable us to perceive the world visually in three dimensions.

FURTHER READING

THE ROLE OF FRAMES OF REFERENCE IN THE DEVELOPMENT OF RESPONSIVENESS TO SHADING. Albert Yonas, Michael Kuskowski and Susan Sternfels in *Child Development,* Vol. 50, No. 2, pages 495–500; June, 1979.

PERCEPTION OF SURFACE CURVATURE AND DIRECTION OF ILLUMINATION FROM PATTERNS OF SHADING. James T. Todd and Ennio Mingolla in *Journal of Experimental Psychology: Human Perception and Performance,* Vol. 9, No. 4, pages 583–595; August, 1983.

PERCEPTION OF SHAPE FROM SHADING. V. S. Ramachandran in *Nature,* Vol. 331, No. 6152, pages 133–166; January 14, 1988.

X-Ray Imaging with Coded Masks

A variant of the pinhole camera that has many apertures arranged in a peculiar pattern can image high-energy X-ray sources, such as plasmas in reactors and black holes in space

by Gerald K. Skinner

How does one make a picture of a distant object that emits only X rays? Telescopes and cameras that rely on ordinary lenses and mirrors are useless. Low-energy X-ray photons, unlike photons of visible light, have enough energy to detach electrons from the atoms of lens material and are thereby stopped by lenses. Higher-energy X-ray photons can pass through a lens, but since they undergo no significant deflection, no focusing can take place. Low-energy X rays can be focused only if they strike multiple reflecting surfaces at grazing angles of incidence. Grazing-incidence reflection, however, becomes less efficient as the energy of the X-ray photons increases, and the angles at which reflection can take place become smaller. A high-energy X-ray telescope would therefore require very large surface areas, and endowing a reflecting surface with the necessary finish and precision is prohibitively expensive. As a result the grazing-incidence imaging technique becomes impractical for X rays of energy greater than about 10,000 electron volts, corresponding to photons with wavelengths shorter than about .12 nanometer (billionth of a meter).

How then can one image objects from which high-energy X rays emanate? Actually the imaging is made possible by a technique that is similar to the one applied in making medical X-ray photographs. To make such photographs a film with an X-ray-sensitive emulsion is placed directly behind a body part that is irradiated from the front by a small "point" source of X rays. The film in effect records the shadow cast by those components of the part that absorb X rays. A somewhat old-fashioned word for the technique is skiagraphy, from the Greek *skia,* meaning shadow.

Now suppose the X rays come not from a point source such as a medical X-ray tube but from an extended source of unknown shape—perhaps a huge cloud of hot intergalactic plasma. Because the extended source can be envisioned as consisting of numerous point sources, the shadow of any intervening object is blurred. The reason is that each component point source causes a slightly different shadow to be cast on the film. Yet if one knows the shape of the intervening object precisely, one can easily predict the form of the shadow it would cast when illuminated by a single point source of X rays. By comparing the shadows produced by all possible combinations of point sources with the actual recorded shadow, it might then be possible to reconstruct the shape of the extended source.

That, in fact, is the operating principle of coded-mask X-ray imaging. The known object between the X-ray source and the detector is a coded mask: an X-ray-opaque plate that has a pattern of holes. The key to this type of imaging lies in the selection of a pattern that enables one to reconstruct an image of the X-ray source from the form of the mask's shadow.

A mask with the simplest pattern possible—a single small hole—is a special example of coded-mask imaging that helps one to visualize how the technique works in general. In essence the system is a pinhole camera [*see illustration on page 86*]. The shadow behind the mask in this case would contain a slightly blurred, inverted image of the illuminating source. Such a pinhole camera will work with high-energy X rays, but it is of little practical use. A sharp image can be obtained only if the pinhole is small, but then few X rays would actually pass through the hole. In an hour only about one X-ray photon from a typical celestial source would pass through a hole one millimeter in diameter in a mask orbiting the earth. Since the background "noise" level (due to cosmic rays) registered by an X-ray detector is far greater than one count per hour, a mask with a single hole in it is not an effective imaging system for cosmic X-ray sources.

An obvious remedy is to perforate the mask several times instead of once, since the number of X rays detected behind the mask is proportional to the number of holes in the mask. Unfortunately each hole produces its own slightly blurred, inverted image, so that a jumble of images is projected onto the detector. How can one recover a single, clear image from such a shadow?

In order to obtain a single recognizable image from the composite one, it is first necessary to divide the field of view into a large number of pixels, or picture elements. Each pixel can then be assigned a "brightness value" corresponding to the flux of X rays that emanate from it. Although each X-ray-emitting pixel gives rise to basically the same shadow of the mask, the shadow generated by one pixel is shifted in relation to the shadow generated by any other. Simple geometry enables one to determine the relative positions of the shadows produced

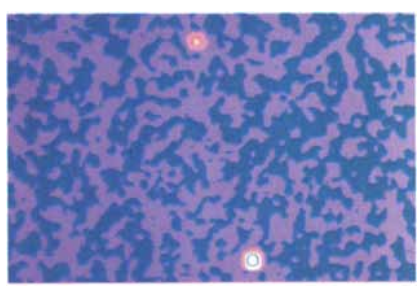
GERALD K. SKINNER is a senior research fellow in the department of space research at the University of Birmingham in England, where he received a B.Sc. in 1965 and a Ph.D. in 1969. He has worked on the detection and imaging of electromagnetic radiation—from infrared to X rays—as applied in the various fields of semiconductor physics, medicine and astrophysics.

by each possible X-ray-emitting pixel.

Since a typical X-ray source can be adequately represented only by many pixels, the shadow that is recorded by the detector is therefore a superposition of a large number of shifted versions of the same basic mask shadow. Nevertheless, one can compare the actual recorded shadow with the shadow that would be generated by each pixel if it were the only source of X rays. A strong similarity (or, to be mathematically precise, a high cross-correlation coefficient) between the shadow that would be generated by a single X-ray-emitting pixel and the recorded shadow indicates that the pixel is in fact a source of a strong flux of X rays. Similarly, a poor match indicates that the particular pixel contributes little or no flux.

By considering each pixel in turn, an image can be built up in this way. Because one has to deal with tens of thousands of detected X-ray photons and perhaps a similar number of pixels, the computing necessary to produce an image is not trivial. Fortunately mathematical techniques that make use of so-called Fourier or Hadamard transforms shorten significantly the necessary computing time and make this kind of imaging feasible.

Of course, it is possible that a mask will produce similar shadows even when it is illuminated by X rays from two quite differ-



DATA from a coded-mask telescope can be presented in "pointillist" form. Every point shows where an X-ray photon struck a rectangular detector. (The cross results from a set of obstructing struts above the detector.) By means of computer processing the data can be made to reveal an image (*above*) of two celestial X-ray sources: GX3+1 and GX5−1.

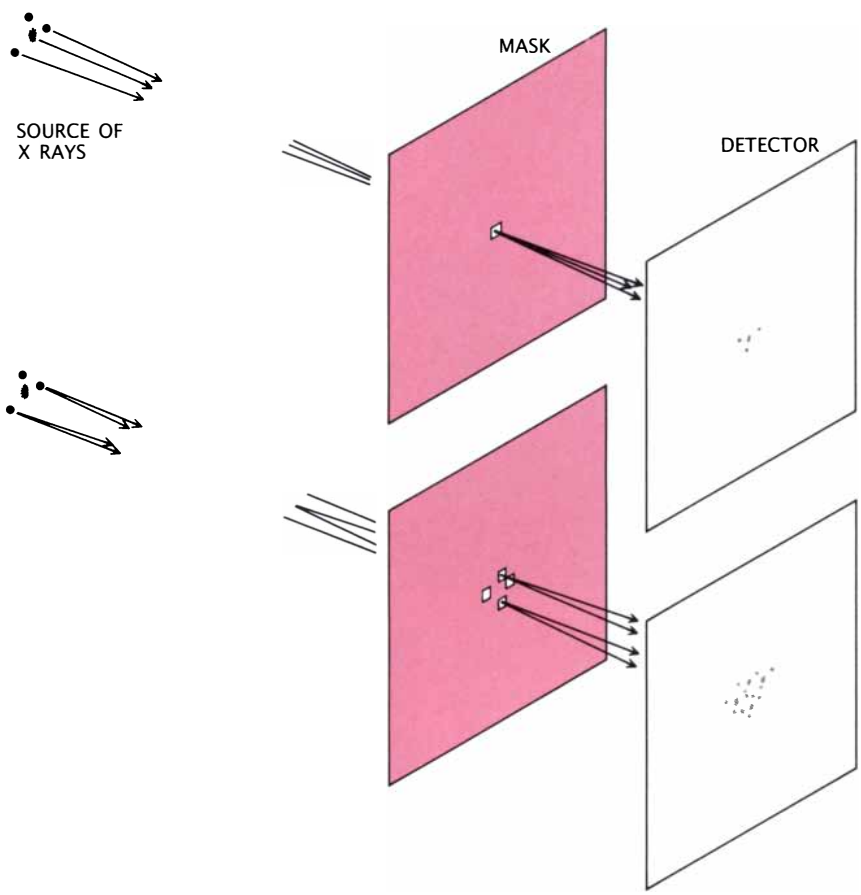
ent pixels. In such a case both pixels would appear bright in the reconstructed image, even though only one of them may actually have emitted the X rays. It is therefore important that a mask's pattern of holes be selected with care, so that the shadow produced when the mask is illuminated by one pixel is as different as possible from the shadow produced by any other pixel. In other words, a shadow cast by the mask must be a poor match to any shifted version of itself.

A regular array of holes would therefore be a rather poor choice: the pattern's translational symmetries could result in shadows that are virtually indistinguishable from shifted versions of themselves, introducing a large degree of ambiguity in the determination of the origin of detected X rays. A much better alternative would be to pepper the mask randomly with holes, and indeed masks with random hole patterns have been applied. Yet a random pattern for the holes is actually not the best choice.

The problem of selecting a hole pattern for a mask in fact resembles familiar problems in other fields of science and engineering. For example, the digital codes for sending information over noisy telephone lines ought to be as different as possible from one another so that if a few bits of a message are corrupted, a totally wrong code is not received. Similarly, a test pattern for checking the alignment of optical systems should be one that is clearly distinguishable from a shifted version of itself.

Optimal solutions to these types of problem are based on mathematical entities called cyclic difference sets, which are best described by an example. Arrange the whole numbers from 1 to 15 in a circle [*see illustration on page 88*], so that counting 15 whole-number steps from 1 brings one back to 1. Among the numbers on the circle the numbers 1, 2, 3, 5, 6, 9 and 11 have a curious relationship to one another. There are three instances in which members of this set are next to each other (1 and 2, 2 and 3, 5 and 6). There are also three instances in which members of the set are separated by one step (1 and 3, 3 and 5, 9 and 11). In fact, on examination one can see that there are always exactly three pairs of members that are separated by any given number of steps on the circle, except for multiples of 15. Similarly, if one replaces the members of the set by holes and rotates the circle by any number of steps, then only three holes will generally coincide with the original position of the holes—not a good match, considering there are seven holes in the set. The exception occurs when one rotates the circle by a number of steps that is a multiple of 15, in which case a perfect match—seven out of seven—results.

These properties are what make the numbers 1, 2, 3, 5, 6, 9 and 11 a "base 15" cyclic difference set. There are other cyclic difference sets that have other bases and that lead to different numbers of matching holes, but all have the same property: as set members are rotated (or shifted) cyclically, each position results in the same poor match among the members—unless one rotates them through one or more complete revolutions.



SIMPLEST CODED-MASK IMAGING SYSTEM (*top*) consists of an X-ray detector placed behind a mask, or X-ray-opaque plate, that has a single small hole in it. The system functions much like a pinhole camera. Only X rays that converge at the hole can strike the detector to imprint a slightly blurred, inverted image of the X-ray source on it. Since the number of X rays that actually pass through the hole is small, the system has little practical value for imaging weak or distant X-ray sources. A mask with several holes (*bottom*) allows more X rays to strike the detector. Unfortunately each hole produces a slightly shifted image of the source, so that computer processing is necessary to decompose the composite image captured by the detector.

To produce a coded-mask pattern from a base-*n* cyclic difference set, a regular grid composed of equally spaced horizontal and vertical lines is constructed. One then numbers from 1 to *n* the squares formed by the intersecting grid lines. In order to ensure that the properties of the cyclic difference set are carried over into two dimensions, however, the numbering must be done diagonally across the grid. Holes are made in those squares that are numbered with a member of the set.

In practice the basic hole pattern is generated by a computer, which then repeats it four or more times and transfers the resulting pattern onto photographic film. A large coded mask can be made easily by projecting the film onto a metal plate coated with a light-sensitive substance. The metal under areas that have been exposed to light (namely where the holes should be) can then be dissolved away with acid. To make sure the remaining metal is opaque to X rays, it is sometimes

necessary to choose rather exotic metals, such as tantalum, for the plate. Alternatively, one can deposit a layer of a good X-ray-absorbing material, such as gold, on the etched plate. This method has been applied to produce masks with more than 32,000 holes.

But how does one come up with a base-*n* cyclic difference set in the first place? The fact is, such sets have to be searched for among all the whole numbers between 0 and *n,* although there are some general rules that help one to find them. In this respect cyclic difference sets are rather like prime numbers. (Indeed, some of the methods for finding them call for first finding a pair of prime numbers.) Various families of cyclic difference sets have been identified, and mask patterns based on them appear strikingly different [*see illustration on page 89*]. Some look like crossword puzzles and others like pieces of op art. Some have a sparse scattering of holes, whereas others have holes in about half of the possible grid positions.

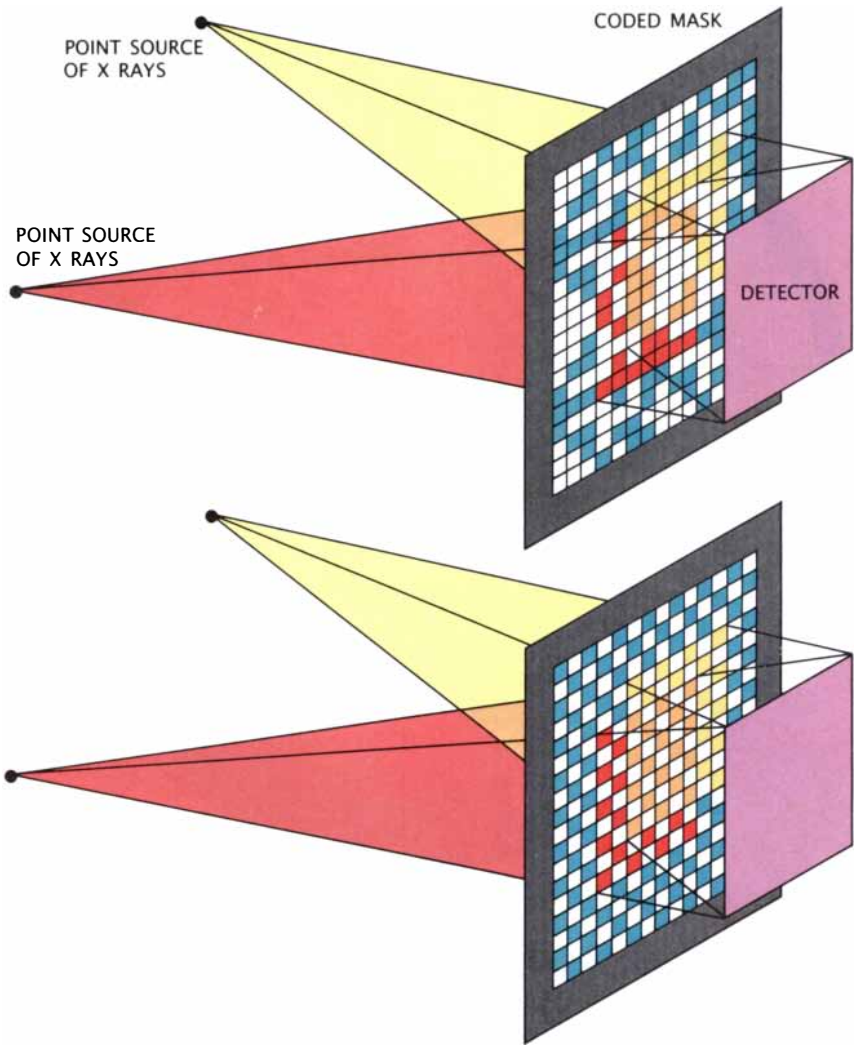
Although films similar to those for medical X-ray photographs could be used to record the X-ray shadow cast by coded masks, they suffer from many disadvantages. They are not sensitive to weak X-ray fluxes, and they cannot distinguish between X rays of differing energy or—for that matter—between X rays and cosmic rays. In addition the film's emulsion quickly saturates: after a grain of the emulsion has been struck by an X-ray photon it does not respond to any further impinging X rays.

For these reasons devices known as gas-filled proportional counters are often used to record the X-ray shadow of a coded mask. (They can also serve to capture the image produced by grazing-incidence X-ray telescopes.) Every time an X-ray photon strikes such a device, electric impulses are produced whose magnitude reveals the energy of the photon. Proportional counters are much more sensitive than photographic films, and their energy-discrimination capability allows one to study the X-ray spectrum of an object. Moreover, the small differences in the arrival times of the pulses at each of the detector's sides can actually be measured, making it possible to determine to a fraction of a millimeter where an individual X-ray photon struck the detector.

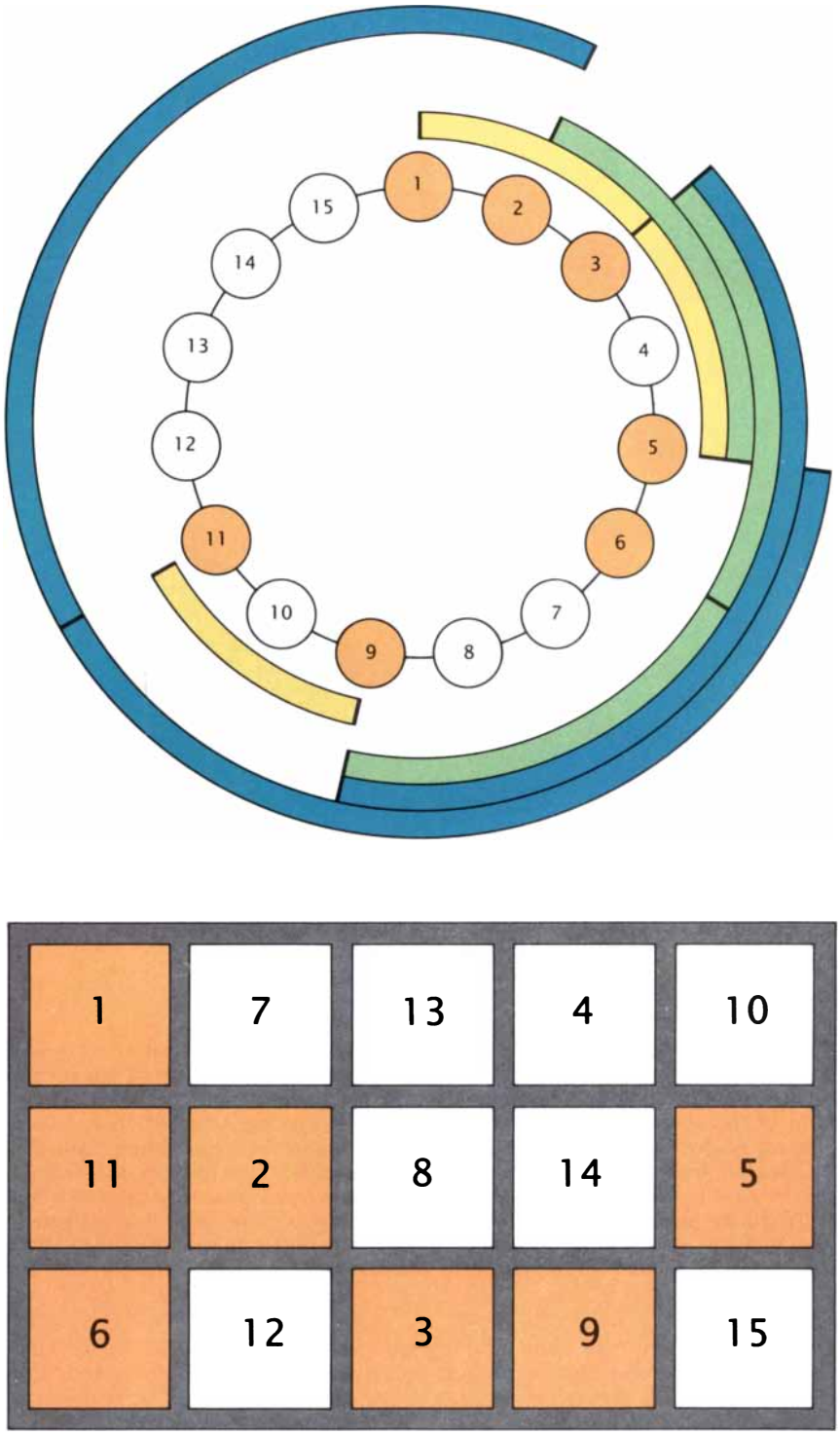
Hence a coded mask, a position-sensitive proportional counter and a data-processing computer are the basic components of a coded-mask imaging system. The mask is placed at an appropriate distance in front of the detector and the assembly is pointed at the source of X rays to be imaged. The pattern of X rays striking the detector is recorded, and if the mask has been properly selected, one can get enough information from the X-ray shadow to reconstruct an image of the source with the computer.

It is interesting to note that the mask-detector assembly acts in some ways like a zoom lens, since varying the distance between the detector and the mask results in different combinations of resolution and field of view. The assembly can also serve both as a telescope for imaging distant X-ray sources and as a camera for forming pictures of nearby objects that are giving off X rays. In either case the data are recorded in the same way. The only difference between the two cases is some slightly varying numerical factors that must be introduced in the computer processing of the data. In essence one focuses a coded-mask imaging system after taking the picture!

One application of coded-mask imaging systems is in fusion research, where plasmas at extremely high temperatures are studied. As the plasma temperature approaches what is necessary for fusion, energy tends to be radiated in the form of X rays. By



TYPICAL CODED-MASK IMAGING SYSTEM (*top*) has a mask with a pattern of holes derived from a "cyclic difference set" (*see illustration on next page*). When such a mask is illuminated by an X-ray source, it casts a shadow that uniquely encodes the form of the source. The reason is that X rays emanating from different points produce shadows on a detector that can be distinguished from each other, in spite of the fact that they are shifted versions of the same basic form. In contrast, shadows cast by a mask that has holes arranged in a regular checkerboard pattern (*bottom*) are virtually identical, regardless of from which point the mask is illuminated. Although only two point sources are shown in this illustration, an extended X-ray source can be thought of as numerous point sources, each producing a shadow.



CYCLIC DIFFERENCE SET of base 15 includes the numbers 1, 2, 3, 5, 6, 9 and 11, which have been marked in a circle of numbers ranging from 1 through 15 (*top*). If the circle is rotated two steps, there will be only three instances where the members of the set coincide with other members in their original positions (*yellow arcs*)—a rather poor match. The same applies if the circle is rotated three steps (*green arcs*) or six steps (*blue arcs*). In fact, there will always be exactly three matches after rotation by any number of steps, except multiples of 15 (in which case there is a perfect match). This is the key property that makes such number sets useful in the manufacture of coded masks: rotating (or shifting) the set members results in a poor match with their original arrangement. To make a coded mask from a base-*n* cyclic difference set, vertical and horizontal lines are drawn to form *n* squares (*bottom*). The squares are numbered diagonally, and a hole is made wherever there is a member of the set. This basic pattern is repeated four or more times to yield a complete mask pattern.

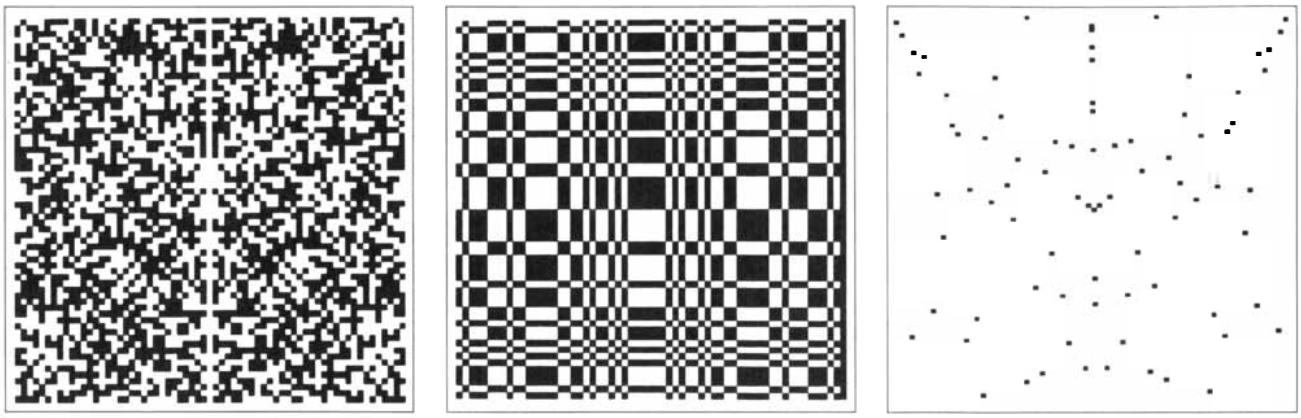
imaging the X rays it is possible to analyze the size and shape of the plasma as well as its temperature variations. Another application is in medicine. Coded-mask cameras can track the X-ray and gamma-ray emissions of drugs that have been labeled with radioactive elements, enabling investigators to see how such substances are absorbed and distributed throughout the body after ingestion.

It is in astronomy, however, that coded-mask X-ray imaging has had the greatest impact. Ironically, astronomers are interested in cosmic X rays because most stars—and indeed most objects in the universe—do not emit X rays to any large extent. The discovery that some astronomical objects do emit copious amounts of X rays came as a surprise. (That discovery was relatively recent, because X-ray astronomy had to await the development of high-altitude balloons, rockets and satellites that could lift imaging systems above the earth's X-ray-absorbing atmosphere.)

When X rays are found to be coming from a particular part of the sky, it means that something rather exceptional is going on there. It implies that there are plasmas at exceedingly high temperatures (at least a few million degrees) or that particles are being accelerated to energies much higher than is usual in the vicinity of normal stars or other common celestial bodies. A list of the types of objects now known to be sources of X rays includes many of the most exotic phenomena in the universe: supernovas, neutron stars, black holes, ultrahot intergalactic plasmas and the nuclei of quasars.

Most images of X-ray sources have been made with grazing-incidence telescopes [see "The Einstein X-Ray Observatory," by Riccardo Giacconi; SCIENTIFIC AMERICAN, February, 1980] and so have generally been limited to sources of low-energy X rays. Before the development of coded-mask imaging systems only the spectrum and intensity variations of high-energy X-ray sources could be studied. Today astronomers can observe some of the most energetic processes known by making X-ray pictures of the sky with coded-mask telescopes.

The feasibility of operating a coded-mask instrument in space was demonstrated in 1976, when a group of investigators (with which I am associated) at the University of Birmingham in England lofted a small coded-mask X-ray telescope into space on a sounding rocket. The first large-scale appli-



BASIC CODED-MASK PATTERNS can appear strikingly different from one another, depending on the particular cyclic difference sets from which they are derived. The ones shown here were constructed by blackening the squares that represent hole locations on grids with a total of 4,095 squares (*left*), 3,599 squares (*middle*) and 6,643 squares (*right*). Because the holes are the same size as the grid spacing, in the first two patterns adjacent holes join to form larger irregular or regular holes.

cation of the technique came in 1985, when two side-by-side coded-mask X-ray telescopes built by our group were flown on the U.S. space shuttle as part of the Spacelab 2 mission.

Each telescope had a different mask pattern that was optimal for somewhat different tasks. The resolving power of a coded-mask telescope, like that of a pinhole camera, depends only on the size of the holes and their distance from the plane in which the shadow is recorded. Both telescopes were about 10 feet long (just fitting within the shuttle's payload bay), but one telescope's mask had small holes (about a tenth of an inch in diameter), resulting in a resolution of three minutes of arc—about a tenth of the diameter of the moon. The mask in the other telescope had holes four times as large and consequently had poorer resolution, but it was more sensitive to diffuse, extended regions of X-ray emission.

In the course of the eight-day mission, images were obtained of many celestial X-ray sources, particularly those identified with clusters of galaxies and the core region of our galaxy. Periodically data from short, sample observations were analyzed on computers in the control room at the Lyndon B. Johnson Space Center in Houston almost immediately after they were recorded. Crude images made from the data were then used to monitor the performance of the instrument. Because of the indirect nature of coded-mask imaging, it was only after the mission was completed, however, that it was possible to appreciate the true wealth of data obtained.

The Birmingham group, working in conjunction with the Laboratory for Space Research in Utrecht, has also built a smaller coded-mask telescope that is one of four instruments for X-ray and gamma-ray astronomy on the Kvant module attached to the Soviet *Mir* space station. The instruments aboard the Kvant module allow large numbers of objects to be observed and their variability in luminosity to be studied. One of the chief targets for the coded-mask telescope has been the supernova in the Large Magellanic Cloud, which appeared fortuitously only a few weeks before the Kvant module was launched. The telescope has been applied to search for the expected emergence of X rays from the supernova as the debris of the explosion expands to reveal radioactive core material.

The images made by the Spacelab 2 and Kvant telescopes have revealed many previously undiscovered X-ray sources and are providing a means of studying the structure of X-ray emission at energies where this kind of observation was not possible before. Particularly interesting is the fact that some X rays come from a position that coincides with the nucleus of our galaxy, where some astronomers think a supermassive black hole may lie. Such an object, a million times more massive than the sun, could very well power some of the extremely energetic processes that have been observed in the region.

In spite of the inherent advantages of coded-mask telescopes, grazing-incidence telescopes are likely to be preferable for imaging sources of low-energy X rays. For one thing, they currently offer the highest resolution. In addition, they can direct the X-radiation from a particular region of the field of view onto a unique part of a small detector; as a result background noise due to cosmic rays and X-radiation from other parts of the field of view can be reduced to a very low level.

Nevertheless, telescopes based on coded masks will be valuable in cases where for one reason or another grazing-incidence telescopes cannot be applied. The commonest reason is the need to image sources of high-energy X rays—or perhaps even low-energy gamma rays. Another reason is the need to image over a wide field of view, as is the case when one is surveying large areas of the sky.

Furthermore, since the resolution of a coded-mask telescope depends on its length, exceedingly high resolution can be potentially achieved merely by making coded-mask telescopes longer. For this reason telescopes hundreds of meters long and even one with a mask and a detector on different space platforms are being considered. The ability to assemble large structures in space may one day allow the development of such telescopes, which would resolve much finer detail in the X-ray and gamma-ray parts of the electromagnetic spectrum than is currently possible.

FURTHER READING

CYCLIC DIFFERENCE SETS. Leonard D. Baumert in *Lecture Notes in Mathematics,* Vol. 182; 1971.

CODED APERTURE IMAGING WITH UNIFORMLY REDUNDANT ARRAYS. E. E. Fenimore and T. M. Cannon in *Applied Optics,* Vol. 17, No. 3, pages 337–347; February 1, 1978.

HADAMARD TRANSFORM OPTICS. Martin Harwit and Neil J. A. Sloane. Academic Press, 1979.

X- AND γ-RAY IMAGING TECHNIQUES. In *Nuclear Instruments & Methods in Physics Research,* Vol. 221, No. 1, pages 1–192; March 15, 1984.

Dr. Atanasoff's Computer

The men who for decades were credited with inventing the first electronic digital computers were not, in fact, first. That honor belongs to a once forgotten physicist named John V. Atanasoff

by Allan R. Mackintosh

History is finally catching up with John V. Atanasoff. After decades in obscurity this 84-year-old retired physics professor is now gaining recognition from computer scientists for something he accomplished almost half a century ago: the invention of the first electronic digital computer. Until very recently, standard histories of the computer routinely credited his feat to others.

Those histories recognized that the computers we know today had their origin in the 1930's and early 1940's, when many complementary and competing attempts were made to automate, accelerate and otherwise eliminate the drudgery of large-scale calculations. In 1932, for instance, Vannevar Bush of the Massachusetts Institute of Technology completed a mechanical computer called the differential analyzer, which did calculus by rotating gears and shafts. Late in the 1930's Konrad Zuse of Germany, George R. Stibitz of the Bell Telephone Laboratories and Howard H. Aiken of Harvard University (in collaboration with the International Business Machines Corporation) independently developed "electromechanical" computers, in which a series of electrically controlled devices known as relays represented numbers. The "on" and "off" positions of the relays stood for the digits 0 and 1 in the binary, or base-2, system. (Unlike the standard decimal, or base-10, system, which represents numbers in terms of the digits 0 through 9, the binary system represents numbers in terms of 0's and 1's.)

ALLAN R. MACKINTOSH has been professor of experimental solid-state physics at the University of Copenhagen since 1970 and director of NORDITA, a Scandinavian institute devoted to the study of theoretical physics, since 1986. He became interested in Atanasoff's story during a visit to Iowa State University in 1983; since then he has spent considerable time studying the early history of electronic computers.

The histories would go on to say that the first electronic computers were invented in the mid-1940's. In contrast to mechanical or electromechanical computers, electronic computers operate primarily by means of such electron devices as vacuum tubes, transistors or, now, microchips; electrons, rather than computer parts, do most of the moving. The first such machine was generally agreed to be the Colossus, which was built by the mathematicians Alan M. Turing and M. H. A. Newman and their colleagues at the Bletchley Research Establishment in England and was operational by 1943. The Colossus helped to decipher the German Enigma code and so decisively affected the course of World War II. The second machine was thought to be the Electronic Numerical Integrator and Computer, or ENIAC, which was built by John W. Mauchly and J. Presper Eckert and their colleagues at the University of Pennsylvania and was operational by 1945.

In reality, between 1937 and 1942—well before either of these impressive and important machines was conceived—Atanasoff had designed and built two smaller electronic computers. The first was a prototype for the larger machine that has come to be known as the Atanasoff-Berry Computer, or ABC. Berry was the late Clifford E. Berry, a graduate student of Atanasoff's and a close collaborator from 1939 to 1942.

The belated recognition of Atanasoff's achievement is not the product of scholarly investigation. Instead it is the incidental result of a lawsuit initiated in 1967 between the Sperry Rand Corporation and Honeywell, Inc. Sperry had bought the patent to the ENIAC and was charging royalties to other manufacturers of electronic computers. Honeywell refused to pay, and so Sperry sued Honeywell; meanwhile Honeywell sued Sperry for violating antitrust regulations and for attempting to enforce an invalid patent.

Honeywell contended that the patent was invalid because in preparing to fight Sperry the company's lawyers had come across a mention of Atanasoff. When they tracked him down, Atanasoff, who had not been privy to the workings of ENIAC, was able to compare that machine with his own. He realized that parts of the ENIAC patent (which covered essentially all aspects of electronic computing) were derived from the ABC and from information he had shared with Mauchly in the early 1940's.

Much impressed by Atanasoff's testimony, Judge Earl R. Larson of the U.S. District Court in Minneapolis concluded on October 19, 1973, that the ENIAC patent was invalid. Mauchly and Eckert, he found, "did not themselves first invent the automatic electronic digital computer, but instead derived that subject matter from one Dr. John Vincent Atanasoff." Both during the trial and after, Mauchly refused to acknowledge that he had learned anything of significance from Atanasoff. Mauchly's widow, Eckert and others also take this view, but in my opinion the court testimony clearly contradicts Mauchly's position.

Larson's decision, which Sperry accepted without appeal, did not immediately bring fame to Atanasoff, in part because the U.S. media were preoccupied with the Watergate scandal that led to the resignation of President Richard M. Nixon. Nevertheless, awareness of Atanasoff's contributions has slowly percolated through the scientific community, and the fact that Atanasoff was the first to design and construct an electronic digital computer is now generally accepted. Much of the credit for that recognition goes to Arthur W. Burks, who was involved in the development of the ENIAC, and to his wife, Alice. The Burkses, respectively professor and a research associ-

ate in the department of electrical engineering and computer science at the University of Michigan, have investigated Atanasoff's work on the ABC thoroughly and have described it—and the patent trial—in an influential article and a recent book.

The path that led to the Atanasoff-Berry Computer essentially began when Atanasoff was working on his doctorate in theoretical physics at the University of Wisconsin at Madison in the late 1920's. His thesis on the electronic structure of helium involved many weeks of laborious computation with a desk calculator and made him long for a more automatic method of computing. Atanasoff's preoccupation with the idea persisted after he earned his degree in 1930 and became an instructor at Iowa State College (later University).

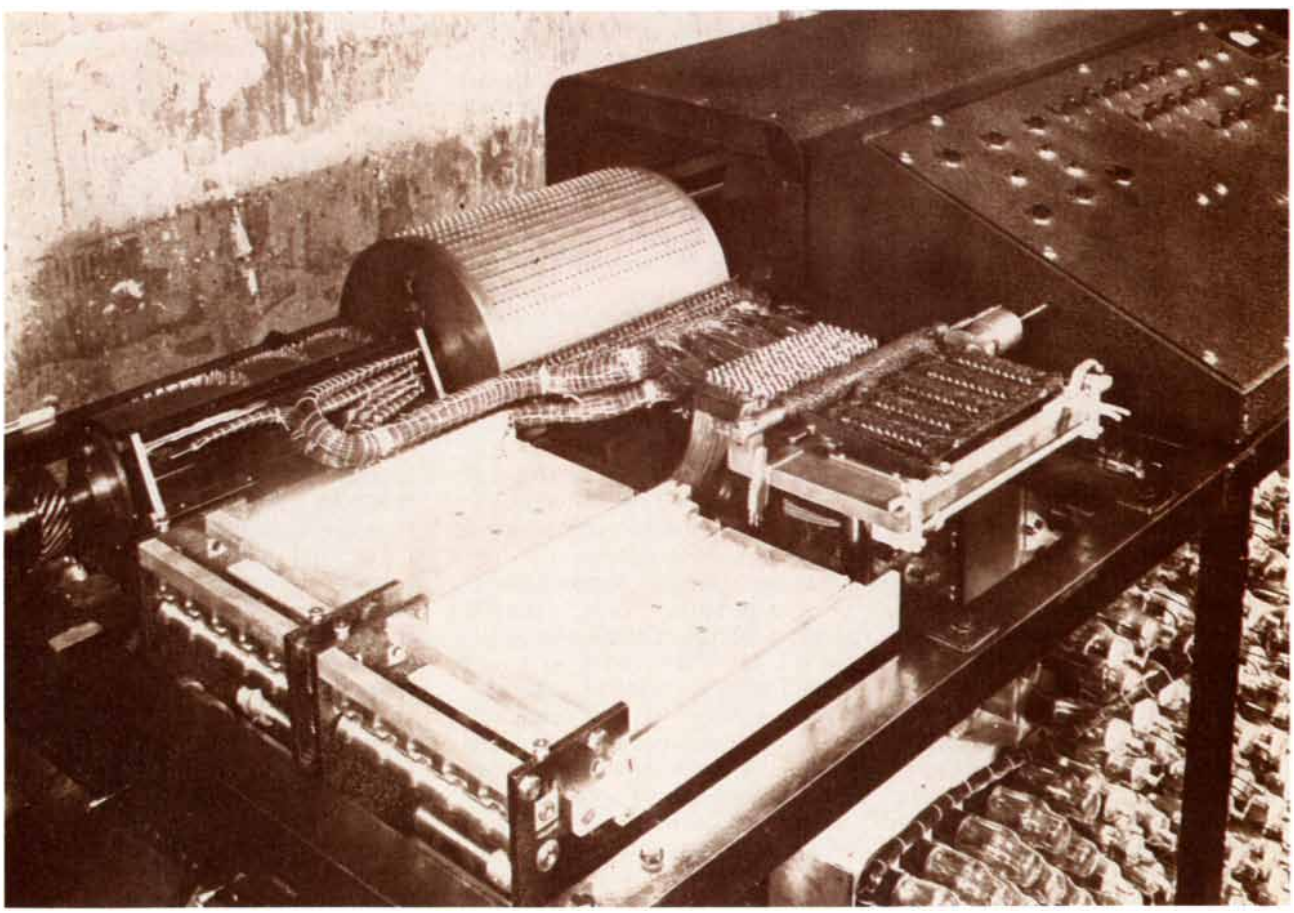
At Iowa he pondered the way to achieve such automation for several years. By the winter of 1937 he had decided on a few general principles. For example, he had determined that the memory function—the storage of data—should be separated from the computational function and that the method of computation should be digital rather than analog: the machine would express numbers as digits rather than by analogy to some physical quantity, such as a distance along the axis of a slide rule. Atanasoff had also toyed with the idea of calculating in terms of bases other than base 10. Nevertheless, his ideas did not seem to "jell," as he put it, and he grew more and more distressed. Then one night of that bleak winter he made several decisive breakthroughs.

The evening had not begun with particular promise. It had, in fact, been so frustrating that he left his laboratory, got into his car and began driving eastward from the college at Ames at high speed, concentrating on his driving to take his mind off his troubles. After several hours he ended up some 200 miles away in the state of Illinois, where he stopped at a brightly lit roadhouse for a drink.

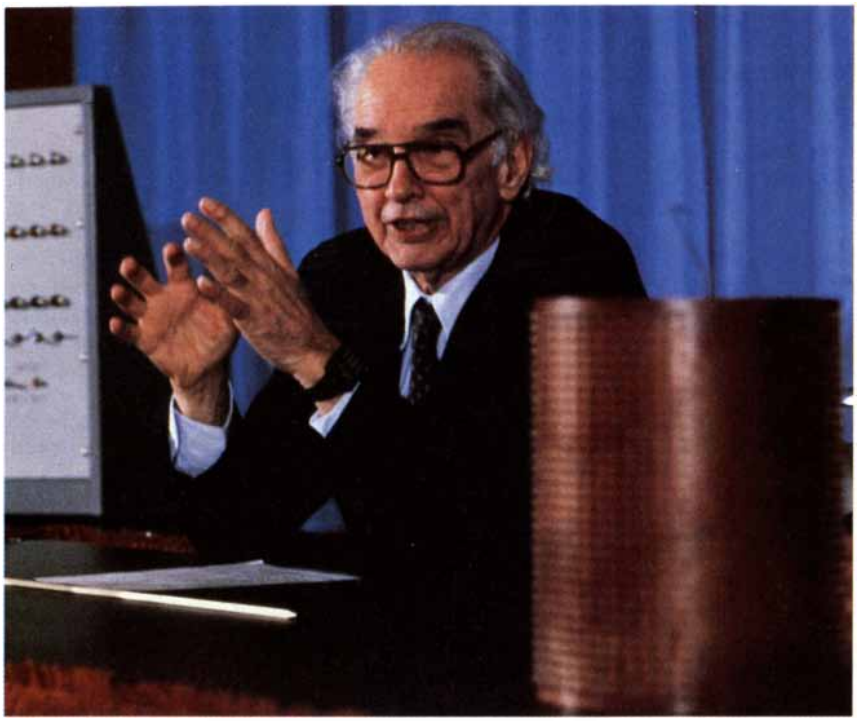
"It was extremely cold and I took off my overcoat," he recalled in trial testimony. "I had a very heavy coat, and hung it up, and sat down and ordered a drink, and as the delivery of the drink was made, I realized I was no longer so nervous and my thoughts turned again to computing machines.

"Now, I don't know why my mind worked then when it had not worked previously, but things seemed to be good and cool and quiet.... I would suspect that I drank two drinks perhaps, and then I realized that thoughts were coming good and I had some positive results."

Positive indeed. Atanasoff resolved to rely on electronic switches (electronic devices that direct the flow of electrical signals), rather than me-



ATANASOFF-BERRY COMPUTER was built between 1937 and 1942 by Atanasoff, then a physics professor at Iowa State College (now Iowa State University), with the help of Clifford E. Berry, a graduate student. The ABC was not the first digital computer ever devised; several earlier machines also manipulated numbers directly instead of representing them as physical quantities, such as the rotation of a pointer. The ABC was, however, the first computer to employ electronics—in the form of vacuum tubes—to carry out computer operations and digital arithmetic; some tubes can be seen at the bottom right. The ABC was also unusual in that the memory and computational elements were separate. The memory consisted of capacitors (devices that store charge) attached to the large drums at the back. The trays at the left are punch-card readers.



ATANASOFF is seen during a celebration of his 80th birthday in 1983 at Iowa State. In the foreground is a memory drum from the ABC, the only major component that survives. Each ring of capacitors on the drum stored one number of up to 50 binary digits, or bits. The drum stored 1,500 bits. Today the total memory in even a simple calculator watch, such as the one on Atanasoff's wrist, can be 10 times that amount.

chanics, to carry out the computer's control and arithmetic functions. In this he was a pioneer. No machine designed for solving complex mathematical problems had been based on electronics before.

He also decided that his digital machine would manipulate binary numbers and would act on those numbers by following rules of logic instead of by direct counting [*see illustration on page 96*]. That same evening Atanasoff also solved a specific problem related to storing numbers in base 2. He had earlier considered employing capacitors, devices that store charge, for the computer's memory. A positive charge on one end of a capacitor could, for instance, represent the number 1, and no charge on that end could stand for 0. The problem was that capacitors have a tendency to lose their charge. Relaxing in the tavern, Atanasoff came up with the idea of regenerating the memory, a process he called jogging. He would restore the charge in a capacitor so that if the capacitor was in, say, the plus state, it would remain that way; it would not change with time or decay to 0.

Having made these decisions, Atanasoff recalled, "sometime late in the evening I got in my car and drove home at a slower rate."

Because modern computers continue to manipulate stored binary digits electronically according to rules of logic and to separate computation and (regenerative) memory, Atanasoff's early decisions are worth examining in more detail. Why, for example, is a digital machine preferable to an analog type for calculation?

Atanasoff's wisdom on that point can best be appreciated by comparing the ABC with Bush's differential analyzer, which was the most advanced scientific computer of the time. In addition to being essentially mechanical, the analyzer was an analog computer: the results were represented by the rotation of a shaft.

Analog computers are suitable for many applications, but in measuring analogous quantities instead of operating on numbers they are subject to an inevitable loss of precision. Atanasoff's digital computer easily attained an accuracy that was 1,000 times greater than was possible with the differential analyzer. Moreover, the precision could readily be increased even further if needed by adding more digits. With analog computers, adding precision is both difficult and extremely expensive. For instance, in order to increase the accuracy of a slide rule by a factor of 10, one would have to increase the length of the rule by the same factor.

Digital computing today is based on the binary system. Clearly Atanasoff was not the only person thinking along these lines—electromechanical computers were often binary—but he was the first to hit on an electronic means of manipulating the binary digits. What does a base-2 number look like? In base 10 each digit in a number represents, from right to left, a given number of 1's, 10's, 100's, 1,000's and so on. Hence the number 237 actually stands for 2 times 10^2, plus 3 times 10^1, plus 7 times 10^0 (any number to the zero power is equal to 1). In base 2 each binary digit, or bit, stands for some number of 1's, 2's (2^1), 4's (2^2), 8's (2^3), 16's (2^4) and so on. Hence the base-10 number 237 would be represented in base 2 as 11101101; counting now from left to right, the number "contains" one unit each of 2^7 (128 in the decimal system), 2^6 (64), 2^5 (32), 2^3 (8), 2^2 (4) and 2^0 (1), and no units of 2^4 or 2^1.

The base-2 system would obviously be impractical for normal use, but because all numbers are represented in terms of 1's and 0's, the system offers the decisive benefit of enabling programmers to represent any number as a series of elements in one of two modes, such as the charged and uncharged states of Atanasoff's capacitors or the "up" and "down" magnetization of regions in a magnetic disk.

Atanasoff decided to store his binary digits in capacitors after considering several alternatives, such as vacuum tubes and ferromagnetic materials (in which the orientations of small magnets can be altered by a magnetic field). He chose capacitors because they were reasonably inexpensive and could send signals to the computational unit without their having to be amplified. This choice, like his solution for recharging the memory devices, continues to influence contemporary computing. Today capacitors are crucial parts of the microchips that form the dynamic memories of modern computers, and Atanasoff's "jogging" is of vital importance to the memory's operation.

Finding a way to preserve memory in capacitors was certainly important, but Atanasoff's greatest achievement was probably the development of a complex electronic switch known as a logic circuit. While he was at the Illinois roadhouse, he had envisioned two memory units, which he called abaci. Then he visualized, as he put it, a "black box"—the logic circuit—into

which would pass the numbers held in memory; on the basis of hard-wired logical rules, the black box would then yield the correct results of an addition or subtraction of the numbers at output terminals.

He decided to build the black box out of vacuum tubes. These would receive signals from the capacitors in the memories, which he named the keyboard abacus and the counter abacus, in analogy respectively to the keys and the movable carriage—the counter—of the mechanical desk calculators popular at the time. The tubes would also receive signals from other capacitors that stored carry-over digits (in the case of addition) or borrowed digits (in the case of subtraction). "Having been taught by a man with a soldering iron," the logic circuit would then select the right answer and replace whatever was in the counter with the result. The tubes would operate on the information so rapidly that they could be enlisted repeatedly to add or subtract the various digits of any two numbers in the abaci. Logic circuits today are stored in tiny chips, which are much faster than vacuum tubes but perform essentially the same functions envisioned by Atanasoff.

What has become of Atanasoff's other major decision, namely to separate memory from processing? His legacy lives here as well. In modern computers, such as the desktop microcomputer, there are three distinct elements: the input-output system, consisting primarily of the keyboard, screen and printer (he decided to have the input and output punched into cards that already existed for use in calculators); the central processing unit, in which the control and processing operations are carried out, and the memory, which has internal and external (disk) components.

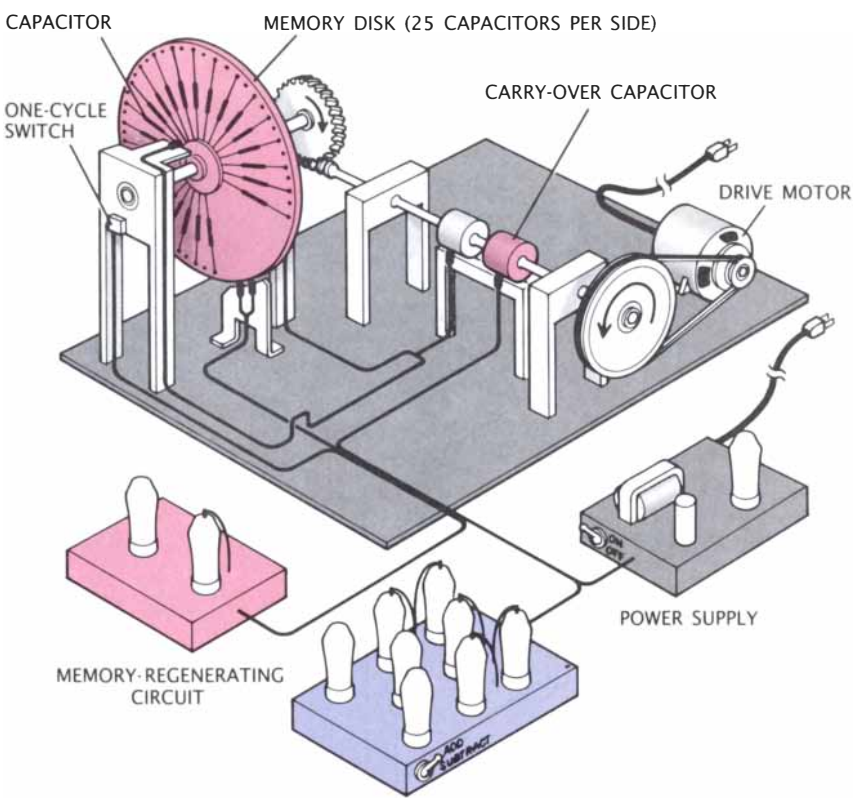
Although Atanasoff was convinced he had found the right principles for electronic computation, he knew that translating these principles into practice would require a prodigious effort. In that effort he received vital assistance from Berry, who was as obsessed as Atanasoff by electronic computing. Atanasoff later recalled that both men were busy, and yet "I do not remember a single instance in which either of us did not have time for the computer; our hearts were really in this adventure."

Their first step was to construct a small prototype to test the essentials of Atanasoff's conception: the electronic logic circuit and the regenerative binary memory. This they did with remarkable speed. The prototype was operational by October, 1939.

It had two memory abaci, mounted on the opposite sides of a plastic (Bakelite) disk. Each abacus consisted of 25 capacitors and hence was able to hold a 25-digit binary number, the equivalent of an eight-digit decimal number. Atanasoff and Berry entered the binary numbers into the abaci by manually charging the capacitors that represented the number 1 and leaving uncharged the capacitors that corresponded to 0. When they pressed a switch, the disk rotated once. As it did so, the electronic logic circuit, which consisted of eight vacuum tubes, read the numbers from the abaci. With the help of a capacitor that held carry-over digits, the circuit then added the numbers and placed the answer in the abacus designated as the counter, for manual reading. At the same time the number in the keyboard abacus was jogged by a regenerating circuit.

The prototype was not a very impressive computer, to be sure: old-fashioned pencil-and-paper computation worked faster. Yet it bears the same relation to electronic computing as, for example, the Wright brothers' airplane bears to aeronautics. By demonstrating the viability of Atanasoff's principles, the prototype opened the path that led to the modern computer.

Atanasoff was now ready to build the ABC proper, which he did between 1939 and 1942. It was designed to do a specific, large-scale computing task that is common in engineering and physics: solving simultaneous linear equations. An example of two such equations is the pair $2x + 5y = 9$ and



PROTOTYPE for the ABC was built in 1939 to test two basic ideas. Atanasoff planned to constantly recharge, or regenerate, the memory capacitors so that they would not lose charge unpredictably. He also planned to calculate by means of logic circuits: sets of vacuum tubes that would add or subtract binary numbers according to logical rules instead of by counting. The prototype was a success. A rotation of the memory disk (*pink wheel*), whose capacitors stored one 25-digit binary number on each side, caused the single logic circuit (*bottom center*) to add or subtract the number on one side of the disk to or from the number on the other side. As the circuit calculated (in the process storing and retrieving carried-over or borrowed digits from one carry-over capacitor), the regenerating circuit (*bottom left*) refreshed the memory.

$x + 2y = 4$, where x and y are the variables, or unknowns. Let us call the first equation *a* and the second one *b*.

As anyone who has studied high school algebra may remember, sets of equations with the same variables can be solved readily by a methodical approach called Gaussian elimination: the addition or subtraction of one equation to or from the other until one coefficient of one variable is equal to 0 and so drops out. In the example here, subtracting *b* from *a* twice reduces the coefficient 2 in $2x$ to 0 and thus results in the equation $y = 1$. When 1 is substituted for y in the original equation *a*, the result is $x = 2$. Note that the process of subtracting *b* from *a* twice amounts to multiplying *b* by 2 and then subtracting it from *a* once; multiplication, after all, is merely multiple additions.

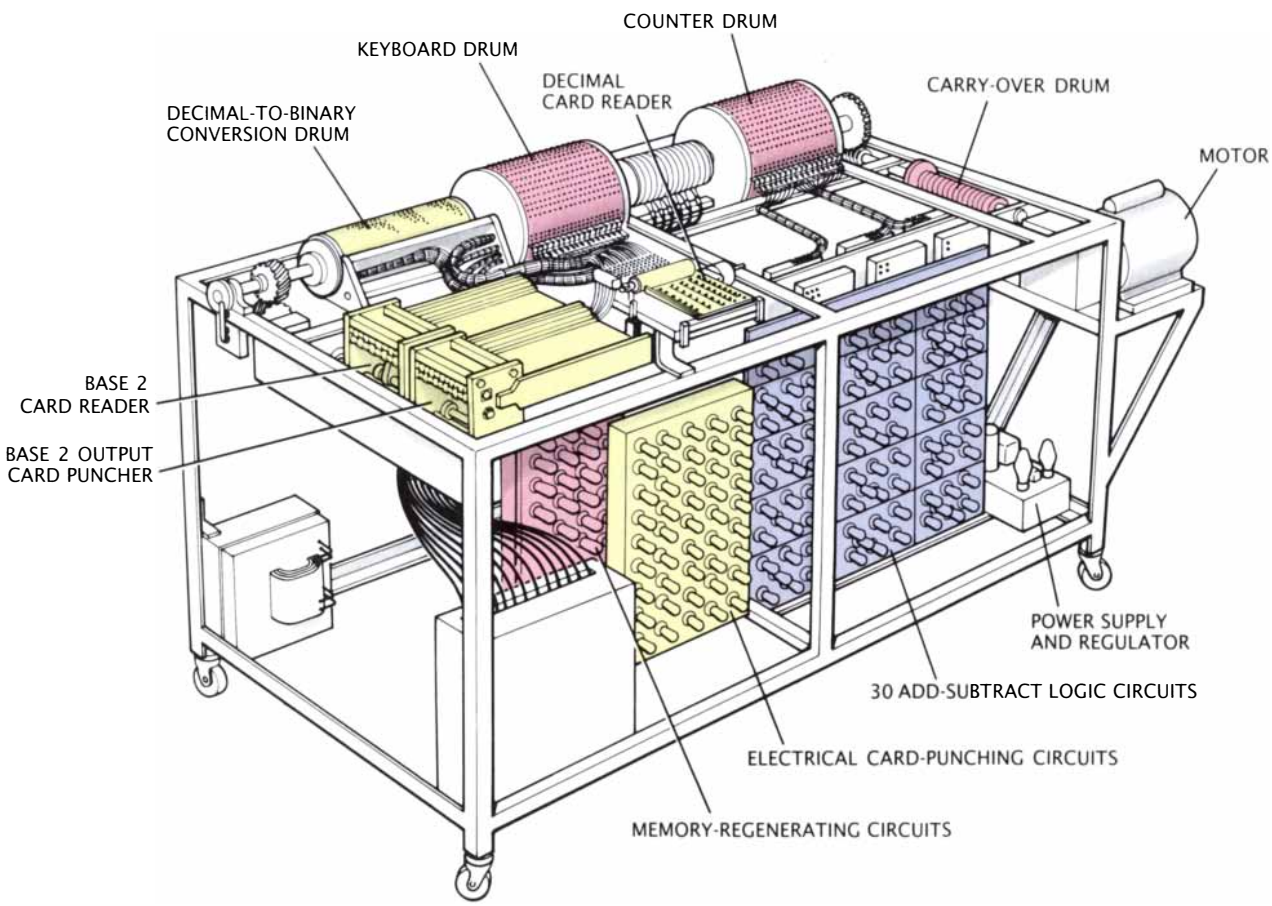
Atanasoff had his sights set on a more complex problem, of course: he wanted to solve n equations with n unknowns—specifically 29 equations with 29 unknowns, x_1 through x_{29}. The solution of such sets of equations follows the example above. As before, one takes two equations—for example $2x_1 + 5x_2 - 3x_3 + 7x_4 + \ldots + 6x_{29} = 9$ and $x_1 + 2x_2 + 4x_3 - 2x_4 + \ldots + 8x_{29} = 4$—and subtracts a multiple of one from the other so that one of the unknowns is eliminated. In order to eliminate x_1, for instance, one would multiply the second equation by 2 and subtract it from the first to produce the result, called the eliminant: $x_2 - 11x_3 + 11x_4 + \ldots - 10x_{29} = 1$.

By repeating this process with different pairs of equations, one can generate 28 equations from which the variable x_1 has been eliminated. Repetition of the procedure with these 28 equations yields 27 equations from which both x_1 and x_2 are missing, and so on until there is only one equation with one unknown. It is then simple to work one's way back up the hierarchy of equations to determine the value of all the variables.

The method is straightforward but clearly involves an enormous amount of arithmetic. Atanasoff estimated, quite realistically, that solving a set of 29 equations with an old-fashioned desk calculator would take about 10 weeks of mind-numbing toil; he estimated that his computer would manage the task in a week or two.

In order to realize his goal of solving many simultaneous equations, Atanasoff placed the keyboard and counter abaci of the ABC on large drums rather than on a disk. Each drum could hold 30 binary numbers that were each specified by up to 50 digits.

The ABC carried out Gaussian elimi-



COMPONENTS OF ABC were designed to enable the machine to solve 29 simultaneous equations, each having 29 variables, x_1 through x_{29}. Such equations can be solved by repeatedly adding one equation to (or subtracting it from) another equation until one variable in the second equation is eliminated. This process is repeated many times to produce the solution: the values of all the variables. To perform such calculations the ABC read the coefficients of the variables (such as the 2 in the term $2x_1$) from prepunched cards, converted them into base 2 and loaded the numbers in one equation into the "keyboard" memory drum and the numbers in the other equation into the "counter" drum. For every rotation of the drums, each of the logic circuits (seven vacuum tubes per circuit) added or subtracted one pair of coefficients, entering the result in the counter. At the same time the memory-regenerating circuits recharged the keyboard capacitors. When the ABC eliminated a designated variable, the machine stored the remaining numbers in the equation on punch cards for later use.

nation for two equations at a time. Their coefficients, which had earlier been punched on cards in decimal form, were converted into base 2 by a specially designed conversion drum and stored in the memory. The coefficients for one equation were loaded onto the counter drum and the coefficients for the second equation went to the keyboard. With each rotation of the drums, which took one second, the logic circuits performed one addition or subtraction on the two sets of coefficients. Specifically, one logic circuit, now consisting of seven tubes, added or subtracted the coefficient of, say, x_1 in the keyboard to or from the coefficient of x_1 in the counter, leaving the sum or difference in the counter. At the same time the other circuits processed the other pairs of coefficients in the same way. (This process, by which a number of identical operations are performed in parallel, is called a vector operation, and a computer carrying out such operations is a vector processor.) Meanwhile still other circuits jogged the keyboard abacus, refreshing the memory.

Later, after multiple subtractions and additions had been performed and a designated coefficient was eliminated, the ABC punched the set of remaining coefficients (the eliminant) on cards in binary form. The cards were then stored until needed in a later step, at which time a binary-card reader transferred the information into the memory. When all the variables had been obtained in binary form, the decimal-card reader operated in reverse to translate the binary data into ordinary numbers.

The punch-card input-output system worked well in preliminary tests, but when it was incorporated into the ABC, an error occurred about once in every 10,000 punching and reading operations. This meant that large systems of equations could not be handled satisfactorily—that is, without extensive recalculation and checking—although small systems could be solved readily. Atanasoff and Berry were still trying to solve this relatively trivial problem when World War II forced them to abandon work on their computer. Berry took a draft-deferred position and Atanasoff joined the U.S. Naval Ordnance Laboratory.

Today the computer they left behind is frequently described as an uncompleted machine. It would be more accurate to characterize it as a functioning but fallible computer, in which the electronic-computing part—the logic circuitry—was a brilliant success. Considering the remarkable speed with which the ABC was designed and constructed, it is safe to assume that the problem with the binary card system would have been solved quickly. Indeed, an input-output system developed decades before by IBM would have been suitable for the purpose (and was later incorporated into the ENIAC). Moreover, by demonstrating the power of his computer, Atanasoff could certainly have obtained financial support to complete the project.

If Atanasoff and Berry had been able to continue, there is little doubt that the ABC would have been fully operational by 1943. Instead it suffered the fate of most aging equipment: it was cannibalized and finally dismantled without Atanasoff's knowledge.

If the ABC was forgotten for so long, how is it that Atanasoff's ideas have influenced modern computing? The answer, of course, lies with Mauchly and his inclusion of Atanasoff's innovations in the ENIAC.

The ENIAC was very different from the ABC. It was the first general-purpose electronic computer, whereas the ABC was designed to be a special-purpose machine. (The ENIAC could be programmed for different problems by altering the configuration of wires plugged into a control panel.) Mauchly and Eckert's machine was much larger than Atanasoff's, with thousands rather than hundreds of vacuum tubes, and it was much faster because its memory was electronic and did not rely on rotating drums. Moreover, the ENIAC calculated by direct counting rather than by logic, and it did so in base 10.

Nevertheless, it is clear that Mauchly and Eckert incorporated Atanasoff's basic elements of electronic digital computing into the ENIAC and a later computer, the EDVAC. Most obviously, the ENIAC and the EDVAC employed electronic switching to control the operation of the computer; the EDVAC also employed logic circuits for arithmetic operations, which were done in base 2, and it made use of regenerative memory. Mauchly also got from Atanasoff the idea that digital electronics would make it possible to build a machine that could do calculus with greater precision and speed than Bush's differential analyzer.

Atanasoff, who by May of 1941 "knew we could build a machine that could do almost anything in the way of computing," decided the ABC could be converted into a digital, electronic differential analyzer after a colleague from M.I.T. told him that workers there were considering incorporating electronics into a new analog version of the analyzer. Atanasoff wrote of the possibility to Mauchly, and the two men discussed it extensively when Mauchly visited Atanasoff for the better part of a week in June, 1941. During that visit Atanasoff also demonstrated the ABC, which was then almost ready to run. Four years later the ENIAC realized Atanasoff's vision.

The ENIAC and the Colossus, which was also programmable, paved the way for the next step in the development of the electronic computer: the incorporation of a program into the memory. This advance in general-purpose devices not only made programming easier and more flexible but also enabled the program to operate differently depending on the results of intermediate steps.

Since the first stored-program computers were introduced in the late 1940's, computers have become faster and more powerful, but their architecture has not changed decisively. There are also echoes of the past in the uses to which some computers are being put. For example, interest in special-purpose computers analogous to the ABC has revived recently, particularly among scientists who have specific problems to solve. In fact, the ABC and a modern vector processor for solving linear equations are astonishingly similar (although the newer machines are enormously faster).

Atanasoff would surely have received earlier recognition for his contributions if he had obtained a patent for his work. As the Burkses point out, he could have laid claim to the concept of electronic digital computation as well as to electronic switching in computers, circuits for logical addition and subtraction, the separation of processing from memory, capacitor-drum memories, memory regeneration, use of the binary number system in electronic computing, modular units, vector processing and clocked control of electronic operations, among other innovations.

It is an understatement to assert that this would have been one of the most important patents ever issued. Unfortunately, because of the confusion created by the war and the ineffectiveness of the people entrusted with the task of obtaining a patent, no patent application for any of Atanasoff's innovations was ever made. For his part, Atanasoff did not take up the patent effort after the war because he was led to believe the ENIAC operated on very different principles from the ABC and that it would be the model for

future computers, rendering valueless any patent for the concepts and devices embodied in the ABC. He was also heavily involved in other projects and, later, in founding his own engineering-research company.

In addition to shedding light on a major achievement in technology, Atanasoff's story provokes a few thoughts about the scientific enterprise. For one thing, the lot of the inventor is not always easy. In spite of much effort, Atanasoff was able to raise only $6,000 for the ABC—whereas half a million dollars was provided to fund the ENIAC because of its military value (among other functions, it churned out firing tables for artillery).

Another thought has to do with the creativity of scientists. Atanasoff's breakthrough on that winter night in 1937 illuminates the creative process with remarkable clarity. He embarked on his project by immersing himself in all aspects of automatic computation. For a long time he grappled with his problem, suffering much frustration and making little obvious progress, but his mind continued to absorb information and to work on it, largely subconsciously. Then, when he was engaged in a completely different activity, the solution came to him.

To the uninitiated Atanasoff's 200-mile drive to the roadhouse might seem to be a particularly inefficient way to get a drink, but he knew perfectly well what he was doing. He appreciated that the mind needs variety and relaxation in order to perform creatively. Having conceived of some fundamental principles, he then allowed "a kind of cognition" to come into play. Such a reliance on intuition may not be in accord with the common perception of scientific investigation as a strictly rational activity, but it is nonetheless an approach followed by many research scientists.

Finally, it is no coincidence that many major advances in technology have been made by scientists: the research endeavor often requires the invention of new tools, and investigators who are deeply absorbed in solving scientific problems are uniquely motivated to rise to the challenge. This connection between science and technology should be understood by anyone who thinks the support of basic research can be restricted without slowing technological progress.

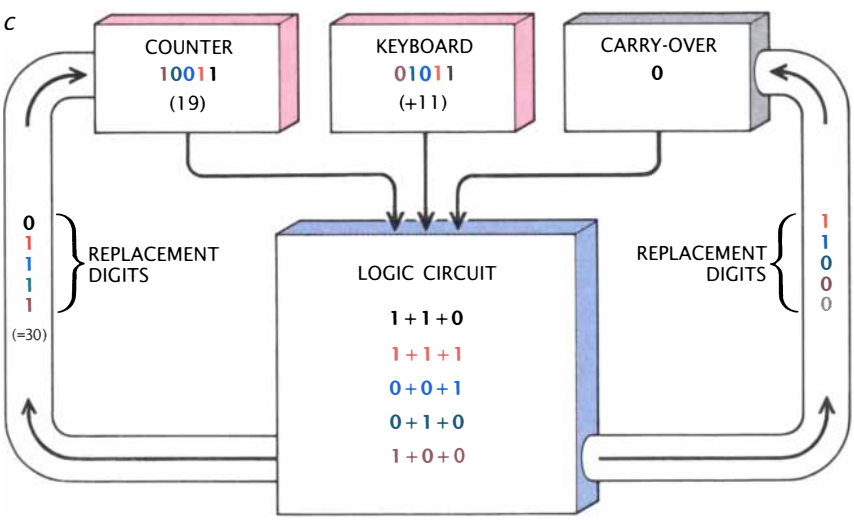
a

	2^4 (16)	2^3 (8)	2^2 (4)	2^1 (2)	2^0 (1)	
COUNTER IN	1	0	0	1	1	(19)
KEYBOARD	0	1	0	1	1	(+11)
COUNTER OUT	1	1	1	1	0	(=30)
CARRY OVER TO NEXT COLUMN	0	0	0	1	1	0

b

INPUT			OUTPUT	
COUNTER	KEYBOARD	CARRY-OVER	COUNTER	CARRY-OVER
0	0	0	0	0
0	0	1	1	0
0	1	0	1	0
0	1	1	0	1
1	0	0	1	0
1	0	1	0	1
1	1	0	0	1
1	1	1	1	1

c



EACH LOGIC CIRCUIT in the ABC added two numbers at a time, such as the ones in the example here (*a*), according to rules laid down in a table (*b*). Assume that the equivalent of the decimal number 19 in the counter and the equivalent of the decimal number 11 in the keyboard had to be added. After the numbers were converted into base 2, which expresses numbers in terms of powers of 2 (such as 2^0, 2^1, 2^2, 2^3, or 1, 2, 4, 8), they would be written as 10011 (16 + 0 + 0 + 2 + 1) and 01011 (0 + 8 + 0 + 2 + 1). The logic circuit would operate on these numbers by first adding the digits in the right-hand column (2^0). To do so it would determine that the configuration of the digits in the counter, keyboard and carry-over memories—1, 1, 0—matched the second-to-last line of the table. (The initial carry-over digit is always 0.) On the basis of the corresponding "output" section, the circuit would then (*c*) send a 0 (*black*) to the counter (where results were registered), replacing the 1 that was there originally. A 1 (*red*) would also be sent to the carry-over memory. The procedure is equivalent to determining by counting that 1 plus 1 equals 2 and that the number 2 in base 2 is written as 10. The circuit would add the digits in successive columns in the same way until the final result was reached.

FURTHER READING

THE ENIAC: FIRST GENERAL-PURPOSE ELECTRONIC COMPUTER. Arthur W. Burks and Alice R. Burks in *Annals of the History of Computing,* Vol. 3, No. 4, pages 310–399; October, 1981.

ADVENT OF ELECTRONIC DIGITAL COMPUTING. John Vincent Atanasoff in *Annals of the History of Computing,* Vol. 6, No. 3, pages 229–282; July, 1984.

THE FIRST ELECTRONIC COMPUTER: THE ATANASOFF STORY. Alice R. Burks and Arthur W. Burks. University of Michigan Press, 1988.

ATANASOFF: FORGOTTEN FATHER OF THE COMPUTER. Clark R. Mollenhoff. Iowa State University Press, 1988.

THE FIRST ELECTRONIC COMPUTER. Allan R. Mackintosh in *Physics Today,* Vol. 40, No. 3, pages 25–32; March, 1988.

SCIENCE AND BUSINESS

Picture Computation

Computer makers vie to build "graphics supercomputers"

Colorful computers, testing for AIDS, sun power, meditation

Above the reception desk at Ardent Computer in Sunnyvale, Calif., hangs a mammoth T-shirt emblazoned with an optimistic sketch of "The World of Supercomputing." Ardent, not surprisingly, occupies most of the foreground. Curiously, a Japanese heavy-equipment manufacturer called Kubota looms in the backdrop atop Mount Fuji. Moreover, Ardent's artists have left out the company's main competitor—Stellar Computer, based in Newton, Mass. Stellar's chief executive officer, J. William Poduska, Sr., does not mind. "We make computers, not T-shirts," he quips.

The picture nonetheless suggests that the landscape of computers used for science and engineering is changing. Both Ardent and Stellar are building computers that exploit what the National Science Foundation is calling visualization—the marriage of high-speed computation and three-dimensional graphics.

While supercomputers have dominated computation in science and engineering for more than 10 years, investigators increasingly have found that the numerical results produced by these fast number crunchers can be as complicated as the initial problems. Graphics workstations were born more than five years ago to give investigators more comprehensible representations of their results. Since the workstations have limited computational abilities, hundreds of investigators often compete for time on a supercomputer to run their calculations; they then dump the results onto graphics workstations. If an algorithm in the model is wrong and the results are incorrect, the investigator must begin again from scratch. Ardent and Stellar—and soon a host of other manufacturers including Silicon Graphics and Apollo Computer—aim to bridge the gap between fast computation and three-dimensional images with their "graphics supercomputers."

Computing at about half the speed of the first Cray supercomputer, the new machines display elegant three-dimensional pictures of everything from molecules to air flowing over the wing of a plane. Consequently they give users the ability to alter complex algorithms while a computation is running and see how such changes affect the calculations and results. The opportunity to watch the intermediate calculations "will open up avenues that people haven't thought of before," predicts James J. Thomas, a principal scientist at the Battelle Pacific Northwest Laboratory in Richland, Wash. Moreover, Thomas adds that he is comfortable with the $100,000 to $200,000 asking price—depending on the number of options—for either Ardent's or Stellar's machine.

Entrepreneurs Allen H. Michels of Ardent and Poduska of Stellar are betting that there is enough demand for visualization to create lucrative business for their start-up companies. Poduska says he looks for more than $18 million in revenues this year; according to Michels, Ardent already has orders for some 40 machines. They predict the marketplace will be worth more than $500 million by 1990.

So far the molecular-modeling community has shown the keenest interest in visualization. The new molecular-science research center at Battelle hopes to buy several graphics supercomputers. Scripps Clinic and Research Foundation in La Jolla has machines from both Ardent and Stellar. Other groups likely to welcome visualization include engineers who are designing mechanical parts, investigators studying fluid flow or military personnel watching real-time simulations of a landscape through the periscope of a tank.

Both Poduska and Michels became intrigued with visualization more than three years ago. During his tenure as chairman of Apollo, Poduska, an engineer by training, foresaw that the emergence of application-specific integrated circuits would enable engineers to pack both computation and graphics power into a relatively inexpensive computer. He first hoped to launch Stellar as an offshoot of Apollo, but the plan did not work out. "How do you convince the shareholders that full efforts are being applied to Apollo, if [its president] has stock in Stellar?" Poduska asks.

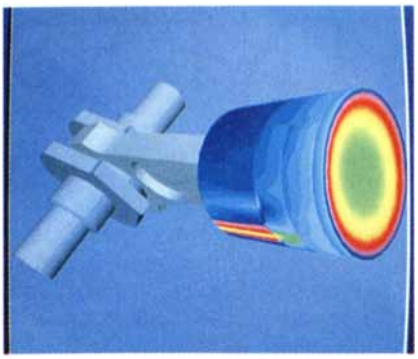
At about the same time, Michels, whom one Ardent staffer describes as the quintessential salesman, itched to leave his position as president of Convergent Technologies and start another venture. Conversations with C. Gordon Bell, father of the VAX workstation, convinced him that there was a growing need—and market—for computers for visualization.

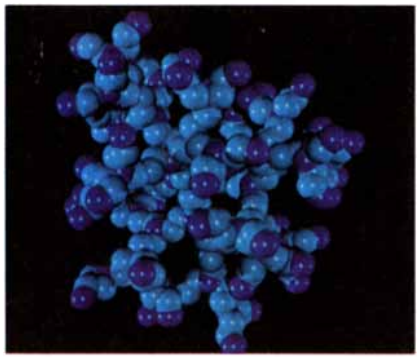
Michels and Poduska were already longtime acquaintances. They kept each other posted on their progress and competed in hiring staff. At one point Michels even asked Poduska to join him. "I live in a different part of the world," Poduska told him. Both tried to lure James H. Clark, chairman of Silicon Graphics in Mountain View, Calif. Instead Clark launched a development project at Silicon Graphics that he says will result in a comparable system by the end of the year. "We wouldn't have gone after it so aggressively if they [Poduska and Michels] hadn't come along," he concedes.

The two men structured their nascent companies along different principles. From the outset at Ardent "there was no question that we'd have our products made in Japan," Michels says. "We needed to manage our support costs and provide the highest possible quality." Ardent found a partner in Kubota, an Osaka-based heavy-equipment manufacturer that was eager to diversify into computers.

Kubota brought a powerful hand to the venture. It spent $70 million building a manufacturing facility strictly for producing Ardent's computer. Since it is difficult to recruit senior-level managers in Japan, Kubota arranged to borrow a director of manufacturing from Oki Electric Industry for five years. (Oki continues to pay his salary.) Industry sources say Kubota gave Ardent some $48 million in venture funding. As a result Ardent designs the software and hardware and Kubota manufactures the complete systems. The deal reportedly gave Kubota 40 percent of Ardent.

Stellar was equally anxious to find a manufacturing affiliate, but not to





With graphics supercomputers, investigators can immediately see how changes in data or algorithms change the structure of the solution. Stellar's computer calculated the piston **(left)*; Ardent's system portrayed the molecule*** **(right).**

trade away stock. It took a more conventional route and agreed to pay Texas Instruments to build and test the printed-circuit boards for Stellar's computer in Johnson City, Tenn. Stellar does the final assembly. Since TI carries the inventory, "it's very good for the balance sheet," Poduska notes.

In spite of the companies' different business approaches, their machines seem much alike. "At this stage the best place to look for the differences is in the blueprints," says Michael Pique, a scientist with Scripps Clinic who has used both systems.

Stellar and Ardent both employ some of the architectural features of supercomputers such as the Cray, teamed with software that automatically looks for ways to process instructions in a parallel way. Both adopted UNIX-based operating systems but are closely watching the newly formed Open Software Foundation, which hopes to build an alternative to UNIX. Ardent chose to employ more off-the-shelf processors than Stellar did but wrote its own software for drawing three-dimensional graphics. Stellar designed all but the most rudimentary components and adopted a de facto standard graphics software.

Nevertheless, what will distinguish the various visualization products will be software rather than raw performance, Thomas notes. "They need to fit into the whole scientific-discovery process." —*Elizabeth Corcoran*

Testing Sales

Will business grow for DNA-probe tests?

Several years ago market analysts were forecasting that DNA-probe tests would be earning hundreds of millions of dollars for biotechnology and pharmaceutical companies by 1990. These days, says Eric B. Rosenbaum, a senior consultant at Arthur D. Little, he has trouble believing the diagnostic tests will even earn $80 million within the next two years. Expectations simply outstripped reality, observes William G. Gerber, a senior director at Cetus in Emeryville, Calif. A DNA probe test for AIDS, however, may show that such tests may be particularly well suited for viral diseases that can lie dormant in cells.

DNA probes are pieces of DNA that match the gene sequences of a particular virus. When a probe is mixed with prepared DNA from a patient, it will bind to any strands bearing the matching sequence—evidence that the patient carries the virus. In order to determine whether there is an infection an investigator has only to check how much material matches.

Since in theory such probes could unfailingly identify diseases, they were predicted to be lucrative business. Mundane problems interfered, however. In practice the tests are complicated and time-consuming to perform. Since the genes under investigation are only a small fraction of the DNA, it is often difficult to get a sample large enough for an accurate test. Even when there is adequate material, investigators typically dye samples with radioactive labels to highlight the results, and technicians do not like to handle radioactive material. Moreover, many of the first probes introduced had a lukewarm reception because they diagnosed diseases that could be adequately identified by other, simpler tests, says Rosemary J. Versteegen, a vice-president at Life Technologies in Gaithersburg, Md.

That scene is changing, albeit slowly. Cetus investigators led by Kary B. Mullis have developed a DNA amplification system based on a technique known as polymerase chain reaction (PCR), which can multiply a minute amount of DNA into a usable sample (see "Supertests" in "Science and the Citizen," February). According to John J. Sninsky, a scientist at Cetus, the reaction produces so much material "it's going to be easy to detect" matching strands without using radioactive dyes. Late last year Cetus and Perkin-Elmer jointly introduced an instrument based on PCR; orders already exceed supplies. Although PCR still does not make using a DNA-probe test easy, for a disease that is not readily detected any other way "people are prepared to go through a lot of hassle," Versteegen observes.

Accordingly Cetus, in collaboration with Eastman Kodak, planned to release a DNA-probe test for the AIDS virus to two California reference laboratories this summer. Instead of measuring antibodies produced by the body in response to the virus, a DNA probe looks for the virus directly. (In some cases patients infected with the AIDS virus have not developed antibodies.) Gerber says the price per test is likely to be between $100 and $200, compared with approximately $30 for an antibody test. Sninsky says that the probe will initially be particularly valuable in detecting early infections, monitoring the progress of an infection, testing babies born to mothers with AIDS and determining how many people become infected after limited exposure to AIDS.

So far the AIDS test seems able to detect "most, if not all," cases of the virus, Sninsky says. Nevertheless, the results are still very preliminary. Until there are more studies of the probe, Sninsky expects the test will only supplement antibody tests. Furthermore, "the antibody test is really simple," he says. "If the test isn't simple, it can't be applied to the blood banks." —*E.C.*

Sunny Prospects?

Photovoltaic technology makes slow but steady progress

It seems everyone has a still better photovoltaic project these days. Companies are planning to build two new U.S. manufacturing facilities, each one able to produce enough photovoltaic cells annually to generate 10 megawatts of electricity. Meanwhile investigators continue to design cells with higher efficiencies. It all follows from the assumption that solar energy will eventually pay off. In some remote locations solar cells already economi-

cally provide power; large solar power plants, however, are research projects.

Even though Federal funding for photovoltaic research has plummeted from about $150 million in 1980 to $40 million last year, industry support has grown to keep research moving apace. Abroad photovoltaic work is growing even faster. West Germany has boosted its funding for photovoltaic research to some 100 million Deutsche marks. Among other projects, it is funding a multimillion-dollar project employing solar-powered water pumps in Africa's Sahel region.

Before photovoltaic power plants become competitive with plants burning fossil fuels, the price of solar cells per peak watt of power must fall by a factor of four, says James L. Brown, president of Solar Cells in Toledo, Ohio. Hoping to lower prices partly through automation, the company plans to begin construction of a 10-megawatt production plant this summer. Chronar in Lawrenceville, N.J., will build the other. By the mid-1990's "photovoltaics for power generation will be a viable option," predicts David E. Carlson, vice-president of the division of Solarex in Newtown, Pa. Solarex hopes to have a fully automated one-megawatt manufacturing facility operating by next year.

All three manufacturers plan to produce solar cells based on thin films of amorphous silicon. (The first solar cells employed thick slabs of single-crystal silicon.) Thin-film cells are widely used for powering calculators, battery chargers and communication devices. A novel application expected to be available later this year is a hot-air exhaust for luxury automobiles. As the sun beats down on a parked car, photovoltaic cells power a fan that pumps out the heated air.

Thin-film cells are cheaper and easier to manufacture than cells relying on crystalline silicon, but they typically generate only about five watts per square foot of cell area compared with 12 watts per square foot for crystalline silicon. Recently, however, ARCO Solar, in Los Angeles, Calif., achieved 10 watts per square foot for a module of thin-film cells made with copper indium diselenide. Investigators also hope to up the efficiency of thin-film cells by stacking two or three films of different materials that are sensitive to different wavelengths of light, in order to take advantage of a larger part of the spectrum. By stacking amorphous silicon and copper indium, for example, ARCO investigators produced 13 watts per square foot at peak sunlight.

The most efficient—and most expensive—devices continue to be "concentrator" cells, which are often made with gallium arsenide. These devices are fitted with lenses that concentrate light to several hundred times its normal intensity and focus it on the cells. Earlier this year the Electric Power Research Institute (EPRI) and Varian Associates, both in Palo Alto, Calif., demonstrated prototype concentrator cells that converted 28 percent of the solar energy striking them into electricity—almost three times as much as current thin-film devices. (According to a Varian manager, the efficiency of a module of manufactured concentrator cells would probably fall closer to 20 percent.)

In spite of the higher efficiencies of concentrator cells, there is little industrial interest in making them for terrestrial applications, says Edgar A. DeMeo, EPRI's program manager. Varian says its large-area gallium arsenide cells (without the lenses) will most likely help to power satellites. EPRI, a nonprofit organization, needs "someone who would like to make these cells and sell them," DeMeo adds. For now, however, most companies believe developing thin-film technologies will bring profits sooner. *—E.C.*

Entrepreneurial Spirit

The Maharishi's university is a high-tech magnet

Fairfield, Iowa, population almost 10,000, boasts one Indian restaurant, two golden domes and about 300 start-up companies. The common factor among all of them is a community of a few thousand transcendental meditators.

Over the past 15 years the two large domes on the campus of the Maharishi International University have drawn meditators to this Middle Western town from all over the U.S. On their arrival the meditators—many of whom were highly qualified in more prosaic pursuits—found tranquillity but few jobs. And so they began an economic renaissance in Fairfield by founding small companies that include computer-software producers, engineering consultants and even an ice-cream manufacturer.

Is this a little offbeat for a town occasionally called the "buckle of the Bible Belt"? That is just what the townspeople of Fairfield thought when the university first started holding classes in the fall of 1974. The apprehensions, however, have "mellowed," says Michael R. Bowers, executive vice-president of the Fairfield area chamber of commerce. Bowers neither meditates nor keeps track of how many of the local businesses are now run by meditators. "They all create jobs and make payrolls," he says.

"Many of us were skeptics about the 'mission' of the university," concedes Harvey Siegelman, Iowa's state economist. But he has been won over. The university quickly gained accreditation and has attracted a total of some 800 students. It professes to be particularly strong in science, in part because science offers "a way to describe consciousness," according to a university representative. As it happens, the father of transcendental meditation, the Maharishi Mahesh Yogi, was originally a physicist in India.

Moreover, the university has become a magnet for new ideas and entrepreneurial zeal, Siegelman says. About six years ago, for example, Daniel W. McGee drove into Fairfield to drop off a friend; McGee was on his way to California to start a company. "But they rolled out the red carpet," and McGee found enough seed capital and engineers to convince him to base Magnetics Research International in Fairfield. McGee, president of the company, has since taken up meditating; he said that about half of his 12 employees also meditate.

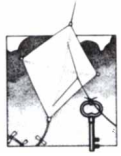
According to Lincoln Norton, president of Corporate Education Resources, a Fairfield software and management consulting company, meditators are an asset in a new venture because they are flexible and "quite cheerful" about changing direction midstream. Norton began meditating in 1966.

Nevertheless, meditating can make it difficult to work standard hours. Most of the community gathers mornings and evenings in the domes. One dome is for men and the other is for women "so you won't get distracted," Norton says. Meditation lasts for a minimum of 20 minutes, although more advanced meditators may continue for more than an hour. As a result "it's tough for a manufacturing plant to hire a meditator," Bowers says. McGee reports his company has flexible working hours to provide time for meditation; he concedes, however, that he only has time for 20 minutes of private meditation at home.

As for the restaurant, business at "A Taste of India" doubled a month ago when owner Robert Schulte decided to switch to a buffet. After all, he notes, between meditating and working "people in this community don't have much time in their day." *—E.C.*

THE AMATEUR SCIENTIST

Some entertaining lessons in optics that may make air travel easier to endure



by Jearl Walker

Reading Elizabeth A. Wood's *Science from Your Airplane Window* is a sure cure for the boredom of air travel (unless, of course, you end up in that terrible row of interior seats). The book is a rich collection of experiments and observations for an airplane passenger. Over the years I have replicated many of Wood's demonstrations and have also scribbled notes about new ones in the margins of her book. This month I explore some of the results.

In particular, an airplane window offers several lessons in optics. The window usually consists of three layers. The outer two are closely spaced and are sealed to the window frame to maintain a comfortable air pressure within the cabin when the airplane flies at high altitude; the third layer protects the sealed layers.

If you are sitting on the shady side of the aircraft, the window may look clear and unblemished, but on the sunny side you may find that the outermost layer is covered with tiny bright scratches. They are visible on that side of the plane because they scatter sunlight to your eyes. Most of the scratches were engraved by rough particles that swept over the window. You see only a fraction of the total number of scratches on a window; other, imperceptible ones are not oriented in such a way as to scatter light to your eyes. If you move your head and change your angle of view, or if the airplane's course changes, you may see a different set of scratches or they may all disappear.

Wood pointed out that the bright scratches form patterns. If you look toward the sun, the bright scratches appear to lie on short sections of concentric circles around the sun. If instead you sit so that the cabin wall blocks your view of the sun, they may appear to lie on straight, parallel lines.

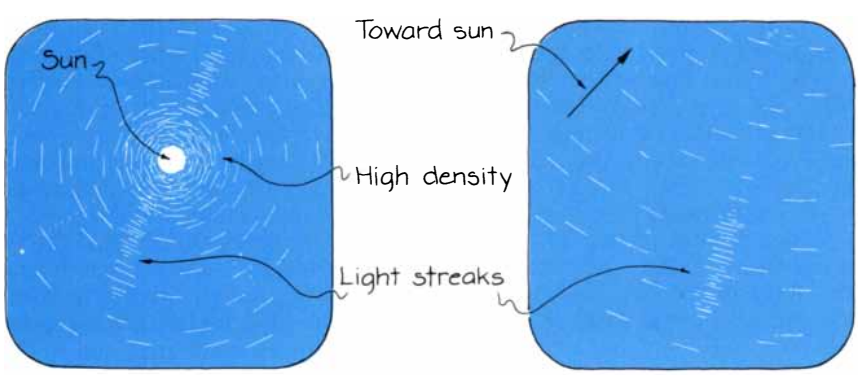
Two features of the scratches puzzled me. When a window is bathed with sunlight, why are only some of the scratches brightly lighted? And why, in the case of those few, is the brightness limited to a small part of the full length of the scratch? After a few flights I figured out that when light scatters from a point along a scratch, it spreads out primarily in a flat fan whose plane is perpendicular to the length of the scratch [*see top illustration on opposite page*]. I estimate that the angle of the fan is small, often less than 30 degrees; the light spreads less than 15 degrees to each side of the scratch. The spread of light out of the plane of the fan is much smaller—something like one degree. Therefore if I see a bright point in the window, my eyes must be in the plane of the fan of light scattered from that point, or within a degree of the plane. The rest of the scratch certainly scatters sunlight, but the light misses my eyes, and so the rest of the scratch is imperceptible—as are all the scratches that send no light whatsoever in my direction.

Suppose the sun is directly off to the side of the aircraft, so that the light rays are initially perpendicular to the window. Any bright scratches to the left and right of the sun are vertical and emit horizontal fans of light that reach your eyes. Bright scratches above and below the sun are horizontal and emit vertical fans of light that also reach you. In addition there are bright scratches at intermediate positions around the sun that happen to be oriented so that they send a fan toward you. Many of the bright scratches lie within 15 degrees of the sun; only a few lie farther out. The density of bright scratches is high close to the sun, because the fans of scattered light are not exactly flat and you receive light from scratches of any orientation.

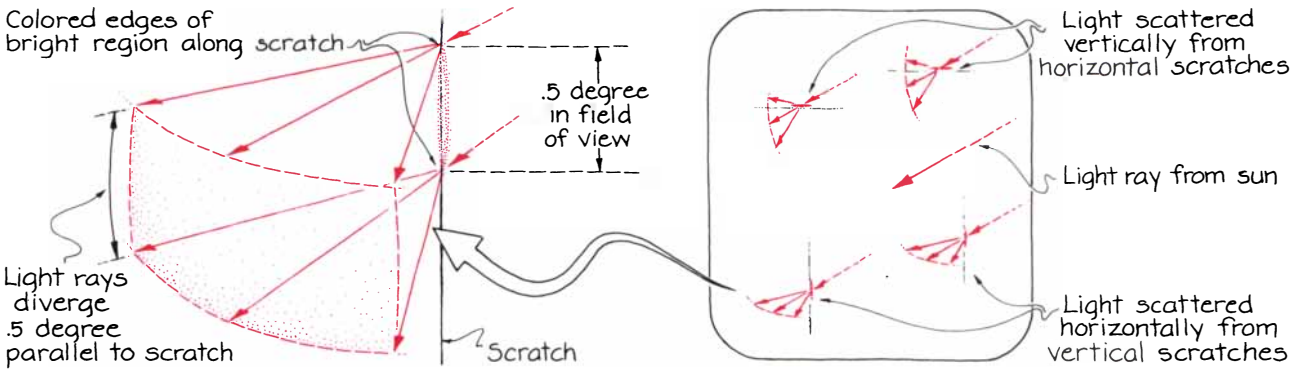
The 15-degree "limit" determines how much of the window has bright scratches. If you sit far from the window, its edges may be within the limit and the entire window will be covered with bright scratches. If you sit next to the window, some of it may be beyond the limit, so that only part of it will have visible scratches. The visibility of scratches also depends on how high the sun is and on the course of the aircraft. For example, if the sun is high, most of the light may be scattered to the floor rather than to you, and the window will appear to be free of scratches.

The arrangement of bright scratches on concentric circles or parallel lines is an illusion. Your brain automatically seeks to impose order on the random pattern of bright spots on the window; provided there are enough bright spots, it brings up to consciousness an illusion that the spots form a geometric pattern. If there are too few bright spots, the illusion disappears and you see the spots as they are: randomly scattered.

When there are abundant bright scratches, I often see long streaks of light that seem to point toward the sun. On close inspection I discovered that the streaks are fans of light, from



Two typical patterns of bright scratches



How scratches scatter light

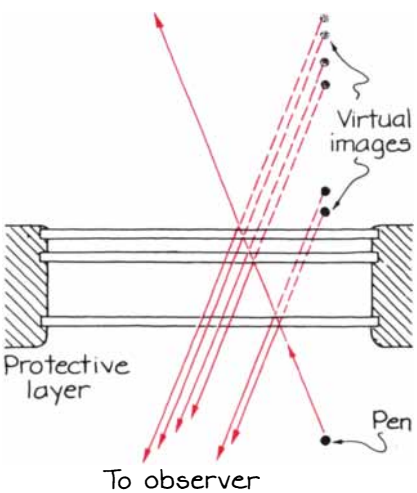
adjacent scratches, that overlap. A close view also revealed that the bright region on a scratch is normally a narrow oval. The short axis of the oval is set by how much of the fan of light is intercepted by the pupils of my eyes. The long axis of the oval is set by the actual shape of the scratch (it may not be straight) and by the fact that rays from the sun are not perfectly parallel but spread by half a degree (because that is the angle the disk of the sun occupies in our field of view). The edges of the oval are often colorful, indicating that the scattering separates the colors in the white sunlight much as a laboratory diffraction grating does.

On a recent flight I spotted another curious optical effect. I was seated on the sunny side of the plane near the front edge of the wing. A jet engine suspended about halfway out on the wing extended forward, beyond the leading edge. Sunlight was reflected from the fuselage to the engine and thence back to me, so that I could see a mirrorlike image of the fuselage on the engine. The outline of the fuselage was distinct in the image, but the row of windows formed a dark band, because the windows reflected sunlight to the engine only weakly. I wondered where I was in the image. Even when I pressed my face up against the window, I reflected too little light to be perceptible. I needed a brighter reflector. I unstrapped my watch and turned its metallic back toward the sun. After I had played with the orientation of the watch for a few minutes, a bright spot appeared in the dark band along the engine: I had found myself.

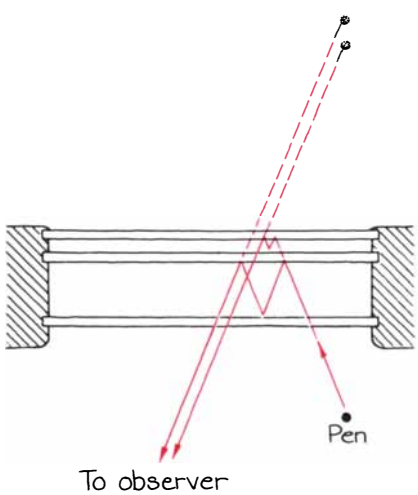
You can also have fun with reflections when you are flying at night, provided you sit next to a window. Turn on the overhead light so that you and the objects around you are illuminated. If you face directly into the window, you will see an image of your face. The edges of the image are fuzzy, because you are actually seeing a composite of three images—one of them from each of the three layers in the window. The image reflected by each layer is as far from the layer as your face is. The farthest layer produces an image that is slightly smaller in your field of view than the image reflected by the closest layer, and the overlap of images yields the fuzzy edges.

Hold a shiny object such as a metal pen out toward the seat in front of you and then look at the window. You will probably see three images of the pen. If you look at the window obliquely rather than directly, you separate the images reflected from each layer. If the view is slanted enough, you will find that each of the three images actually consists of a pair of overlapping images. In each pair the nearer image is a reflection from the front (the near side) of a layer and the farther image is a reflection from the back (the far side) of the layer.

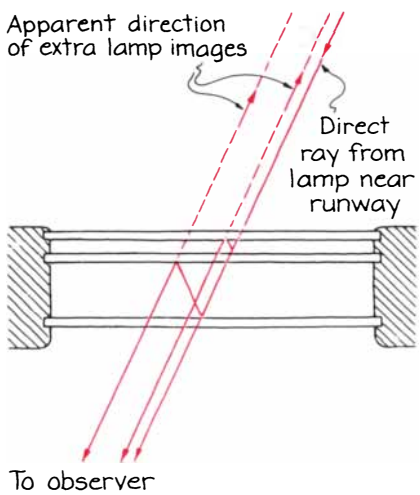
The reflections giving rise to the various images are indicated in the illustration at the left below. Each reflected ray is mentally extrapolated backward, so that it appears to have originated beyond the window. (For clarity the illustration has been simplified by assuming that the observer is far from the window; a single ray from the pen is responsible for the various reflections and images. Actually the



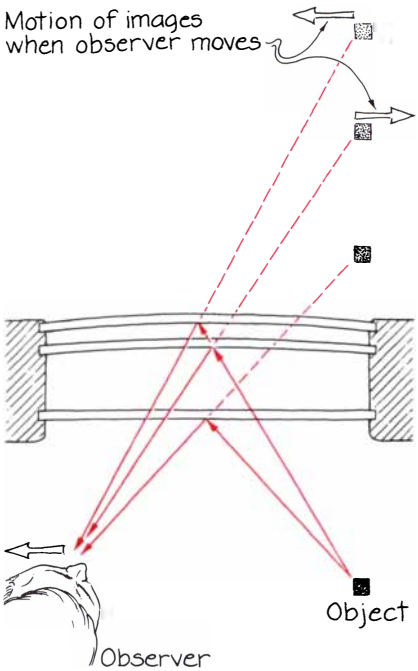
Images of a pen seen in a window



Images due to multiple reflections



Multiple images of a runway light



Reflections from curved windowpanes

observer is near the window and a slightly different ray from the pen is needed for each reflection; all the reflected rays converge at the observer's eyes. I shall ignore that detail here.)

The images are dim because every time the light reaches a surface (either the front or the back of a layer) only a small fraction of the light is reflected; the rest of it continues to travel outward. The dimness of the images makes them hard to make out during daylight, when a flood of light comes to your eyes from outside the plane.

If the night is dark and if the light reflecting off the pen toward the window is particularly bright while the cabin is otherwise dark, you may see even more images in the window. They are caused by multiple reflections between layers or even from the front and back surfaces of a single layer. For example, a ray from the pen can pass through the first layer and be reflected from the middle layer, be bounced back by the first layer and then reflected again from the middle layer before it reaches you [*see middle illustration at bottom of preceding page*]. The bounced light creates another image of the pen, but this image is quite dim, because the three reflections leave a final ray of low intensity. You may notice other images in a window. Can you figure out how they develop from multiple reflections?

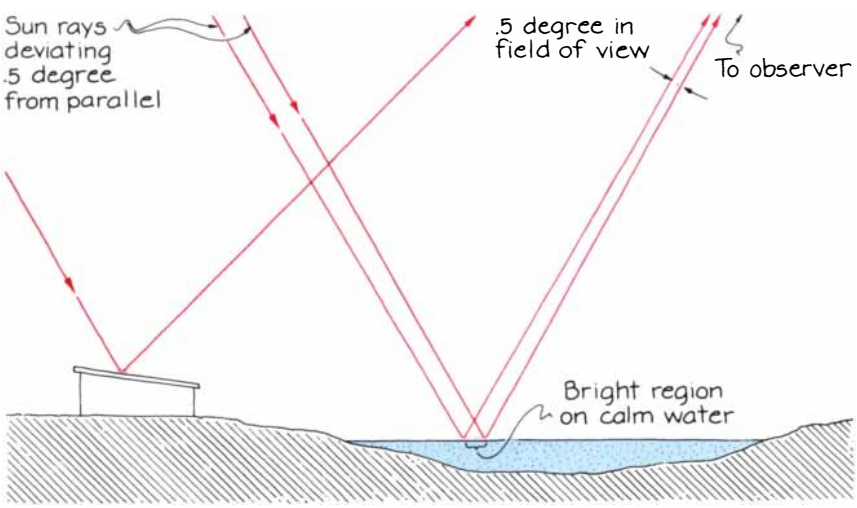
I can see images arising from multiple reflections better when the aircraft is waiting for takeoff at night. I switch off the overhead light and examine a runway lamp either ahead of me or to the rear. Trailing off to one side of my direct view of the runway lamp there are images that must have been generated by multiple reflections from the window layers or from the front and back surfaces of a single layer [*see illustration at bottom right on preceding page*]. A runway lamp that is directly off to the side of the aircraft does not give such a display. The light from it does undergo multiple reflections, but the resulting images overlap and cannot be distinguished. When the light from the lamp reaches the window along a slanted path, the images due to multiple reflections between layers are separated enough to be distinguished; the images due to multiple reflections within a single layer are more tightly spaced and are harder to distinguish.

The image that lies closest to the direct view of the lamp results from a double reflection between the outer two layers of the window. The light passes through the outermost layer, is reflected from the middle layer and then back to the outermost layer before being reflected to my eyes. The next side image is caused by light that is reflected once from the inner layer and then once from the middle layer. Near by, but slightly farther from the direct view of the lamp, is an image due to light that is reflected from the inner layer and then the outermost layer. Other images may also be present, but they are dimmer, because they involve four or more reflections from the layers. Their positions depend on both the angle at which you peer into the window and the angle at which the lamp's light reaches the window.

On a night flight recently I was seated on the right side of the plane in an oddly configured row where there was extra space between the window and my window seat. The space allowed me to peer back into the window just behind me. Before takeoff I sketched how the window yielded three images of an armrest of the seat behind mine. When I moved my head forward, the images maintained their relative spacing. Later, at high altitude, I again examined the reflections. This time their spacing was different. Also, when I moved my head forward, the outer two images moved in opposite directions: the outermost image unexpectedly moved forward and the middle one moved to the rear.

I suspected the reason for the different spacing of the images and their odd movements was that the sealed layers in the window were no longer flat, having flexed outward as the air pressure decreased outside the aircraft. After making a few sketches I understood the images. The outermost one came from light rays that happened to be reflected from just to the left of the center of the outermost layer. When I moved my head forward, I intercepted a different set of rays, which were reflected from farther to the left of center. If the layer had been flat, both sets of rays would have seemed to originate at the same place outside the window, and the image I perceived there would have been stationary. Because the layer was curved, however, the second set of rays appeared to originate to the left of the first set, and so when I moved, the image shifted to the left with me.

The middle image came from rays reflected from the middle layer. These rays were reflected from just to the right of the center of the layer. When



The reflection of the sun's rays from a roof or a pond on the ground

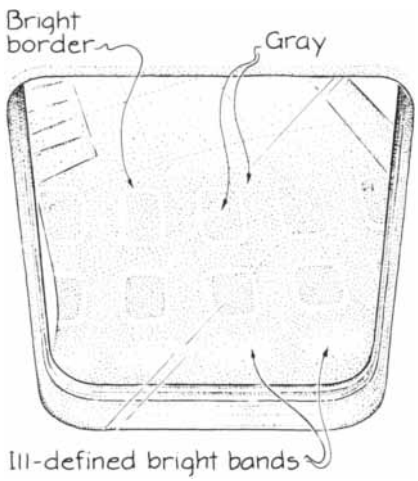
I moved forward, I intercepted rays reflected from farther to the right of center, making the image shift to the right—in the direction opposite to that of the outermost image.

When the plane landed, I again studied the movement of the reflections when I moved my head forward. This time the outermost and innermost images maintained their relative spacing, indicating that the outermost layer was again flat. The middle image still moved in its peculiar way, though: apparently the middle layer of the window had not yet adjusted to the return to normal air pressure. On later night flights I found that the curvature of the layers could also be detected when I held a metal pen in front of me. When the reflections were near an edge of the outer layers, the images of the pen were curved.

If you fly on a sunny day and on the sunny side of the aircraft, watch the reflections of the sun from water surfaces below. The spectacle is best when the aircraft is at a low-to-moderate altitude and when there are no waves on ponds or lakes. A flat water surface is a mirror that reflects a bright image of the sun. If the body of water is large enough, the image occupies the same half-degree angle in your field of view as the sun would in a direct view; a pond that is large enough to reflect a complete image when the plane is low may give only a partial image at higher elevations.

When there are a number of water surfaces below you—ponds, canals or even swimming pools—the sun's image plays hide-and-seek, skipping between water surfaces but always constrained to stay along a path set by the flight of the aircraft; you may appear to be moving over scattered pieces of a shattered mirror. Other shiny surfaces, such as the slanted metal roof of a shed or, if you are low enough, the windows of cars in a parking lot, send up brief flashes of light, which may be well off the sun's path over flat water. Even the bumpers on the cars or the metal traffic signs along roadways enter into play. I was surprised by the reflections from signs, which would not appear to be oriented correctly to reflect the sun upward. Apparently the sun is so bright that even the top edge of a sign can contribute a noticeable albeit fleeting reflection.

Here is another reflection observation, which I shall leave for you to figure out. While flying on the sunny side of the aircraft with the sun above me, I noticed two rows of rectangles on the wing. Each rectangle was marked by a bright border. The rectangles in the farther row were longer than those in the nearer row. Once, when the pilot tilted the wing upward, the rows slid toward the fuselage and a third row appeared just beyond the second one. Obviously the rectangles were images of the windows. The puzzle is: Why were the images not in a single row?



Puzzling images seen on the wing

Air travel provides some lessons in high-pressure physics as well as in optics. Although airliners have pressurized cabins, the cabin pressure is always less than the air pressure on the ground. One way to monitor the change in air pressure in flight is to examine one of those sealed, plastic containers of gelatinous salad dressing that are served with dinner. When the plane is on the ground, the flexible top of the container bulges inward, but when the aircraft is at high altitude, the top bulges out, because the cabin pressure is less than the air pressure inside the container.

If you shake the container to mix its contents, you coat the inside of the top of the container with salad dressing. When you then open the container by peeling off the top (if you can find the peelable edge), the sudden exposure of the interior to the low pressure in the cabin will blow the gooey coating outward. To avoid squirting salad dressing into your lap, take care to aim the opening toward the salad.

FURTHER READING

EFFECT OF ENVIRONMENTAL CHANGES ON THE GHOSTING OF DISTANT OBJECTS IN TWIN-GLAZED WINDOWS. W. Swindell in *Applied Optics*, Vol. 11, No. 9, pages 2033-2036; September, 1972.

SCIENCE FROM YOUR AIRPLANE WINDOW. Elizabeth A. Wood. Dover Publications, Inc., 1975.

SCIENTIFIC AMERICAN

CORRESPONDENCE

Offprints of more than 1,000 selected articles from earlier issues of this magazine, listed in an annual catalogue, are available at $1.25 each. Correspondence, orders and requests for the catalogue should be addressed to W. H. Freeman and Company, 4419 West 1980 South, Salt Lake City, Utah 84104. Offprints adopted for classroom use may be ordered direct or through a college bookstore. Sets of 10 or more Offprints are collated by the publisher and are delivered as sets to bookstores.

Photocopying rights are hereby granted by Scientific American, Inc., to libraries and others registered with the Copyright Clearance Center (CCC) to photocopy articles in this issue of SCIENTIFIC AMERICAN for the flat fee of $1.25 per copy of each article or any part thereof. Such clearance does not extend to the photocopying of articles for promotion or other commercial purposes. Correspondence and payment should be addressed to Copyright Clearance Center, Inc., 21 Congress Street, Salem, Mass. 01970. Specify CCC Reference Number ISSN 0036-8733/88. $1.25 + 0.00.

Editorial correspondence should be addressed to The Editors, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 10017. Manuscripts are submitted at the authors' risk and will not be returned unless accompanied by postage.

Advertising correspondence should be addressed to Advertising Manager, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 10017.

Address subscription correspondence to Subscription Manager, SCIENTIFIC AMERICAN, P.O. Box 953, Farmingdale, N.Y. 11737. The date of the last issue on subscriptions appears at the right-hand corner of each month's mailing label. For change of address notify us at least four weeks in advance. Please send your old address (if convenient, on a mailing label of a recent issue) as well as the new one.

Name

New Address

Street

City

State and ZIP

Old Address

Street

City

State and ZIP

COMPUTER RECREATIONS

The hodgepodge machine makes waves

by A. K. Dewdney

Cellular automatons, computer models based on arrays of multivalued cells, have spread like a wave through physics, mathematics and other sciences. Now a new cellular automaton has literally been making waves of its own. Called the hodgepodge machine by its designers, it imitates chemical reactions with a precision rarely seen in other models.

The reactions the hodgepodge machine simulates take place in excitable chemical mediums: two or more compounds that can dissociate and recombine in the presence of a catalyst. If the chemical states of the reactants have different colors, wavelike structures can be seen that propagate along simple or intricate frontiers in endless pursuit of an elusive equilibrium.

Does the automaton itself serve as an adequate physical explanation for the waves observed in actual reactions? This question now occupies the hodgepodge machine's creators, Martin Gerhardt and Heike Schuster of the University of Bielefeld in West Germany, along with an increasing number of colleagues at other universities.

A cellular automaton can be thought of as an infinite grid of square cells that advance through time in step with discrete ticks of an imaginary clock. At any given tick each cell is in one of a finite number of states. The state of a cell at tick $t+1$ depends in a fairly simple way on the states of the cells in its immediate neighborhood at the previous tick, t. The dependence is expressed in a set of rules that apply equally to all the cells in the grid. By applying the rules each time the clock ticks, an arbitrary initial configuration of states among the cells can be made to change and thus evolve with time. In some cases extraordinary patterns develop, prompting observers to believe that given the right initial configuration a cellular automaton could produce something capable of organizing itself, growing and reproducing—in short, something "living."

The cellular automaton best known to readers is probably the famous game of Life invented in the 1960's by the mathematician John Horton Conway of the University of Cambridge. In Life each cell has only two possible states: alive and dead. The rules of Life are very simple. If a cell is dead at time t, it will come to life at time $t+1$ if exactly three of its neighbors are alive at time t. If a cell is alive at time t, it will die at time $t+1$ if fewer than two or more than three of its neighbors are alive at time t. These two rules are sufficient for the Life cellular automaton to display an amazing variety of behavior that depends entirely on the configuration of dead and alive cells with which one starts [see "Computer Recreations," SCIENTIFIC AMERICAN, May, 1985, and February, 1987].

The hodgepodge machine is not one cellular automaton but many. One chooses a particular version by specifying a number of parameters such as the number of states. If there are $n+1$ states, each possible state of a cell can be represented by a number between 0 and n. Gerhardt and Schuster extend Conway's metaphor to describe the states of the cells in their machine. A cell in state 0 is said to be "healthy" and a cell in state n is said to be "ill." All states in between exhibit a degree of "infection" corresponding to their state number; the closer a cell's state number gets to n, the more infected the cell becomes. The hodgepodge machine selectively applies one of three rules to each cell, depending on whether it is healthy, ill or infected.

If the cell is healthy (that is to say, in the 0 state), at the next tick of the clock it will have a new state that depends on the number of infected cells, A, and the number of ill cells, B, currently in its neighborhood and on two parameters labeled $k1$ and $k2$. To be specific, the state of the cell at time $t+1$ is given by the following formula:

$$[A/k1] + [B/k2].$$

A pair of square brackets designates a rounding-down process applied to the fraction it contains. If, for example, $A/k1$ happens to equal 2.725, the square brackets reduce that number to 2. If the formula happens to yield a 0, the cell will of course remain healthy—at least for the time being.

If the cell is infected, its condition generally worsens with time. Its state at time $t+1$ is the sum of two numbers: the degree of infection in the cell's neighborhood at time t and an unvarying quantity, g, that governs how quickly infection tends to spread among the cells. The degree of infection is calculated by dividing S, the sum of the state numbers of the cell and of its neighbors, by A, the number of infected neighbors. A cell in an infected state at time t therefore takes on at time $t+1$ a state given by the formula

$$[S/A] + g.$$

The infected cell cannot get "sicker" than n, however. If it happens that the number given by the formula exceeds n, then n is taken to be the new state of the cell.

Finally, if the cell is ill (in state n) at time t, it miraculously becomes healthy (takes on a state of 0) at $t+1$.

In addition to those three rules a definition of what constitutes a cell's "neighborhood" is necessary. Two types of neighborhood have historically been used in cellular automatons: the von Neumann neighborhood and the Moore neighborhood. The von Neumann neighborhood of a particular cell consists of the four cells that share the cell's edges. The Moore neighborhood of a particular cell includes the cells in the von Neumann neighborhood and also the four cells that just touch the cell's corners—a total of eight cells. Given the three rules and the definition of a cell's neighborhood, the Gerhardt-Schuster cellular automaton is completely defined by specifying the values of four parameters: n, the number of states

minus 1; $k1$ and $k2$, the "weighting" parameters for healthy cells, and g, the speed of infection.

A sample experiment done by Gerhardt and Schuster on a 20-by-20 grid using von Neumann neighborhoods reveals the typical behavior of hodgepodge machines. (Cells at the edge of the grid abide by the same rules that prevail elsewhere in the cellular automaton; they just have fewer cells in their neighborhood.) The parameters n, $k1$ and $k2$ were fixed respectively to the values of 100, 2 and 3. Four types of behavior emerged at different values of the parameter g. In a typical trial run Gerhardt and Schuster gave the 400 cells in the 20-by-20 grid a random initial configuration of states, specified a value of g and let the hodgepodge machine loose for 10,000 computational cycles. Because one-dimensional data are easier to analyze than two-dimensional images, Gerhardt and Schuster recorded only the number of infected cells at each cycle in order to present their results in graphs like those on the next page.

Not much happened to this hodgepodge machine at low g values. Apart from a few initial fluctuations, activity among the cells tended to die out; the cells became boringly and everlastingly healthy. But as g was increased, strange things began to happen. To begin with, most of the cells became infected and remained so, although there were irregular and random appearances of healthy cells. Gerhardt and Schuster labeled this type of behavior Type 1.

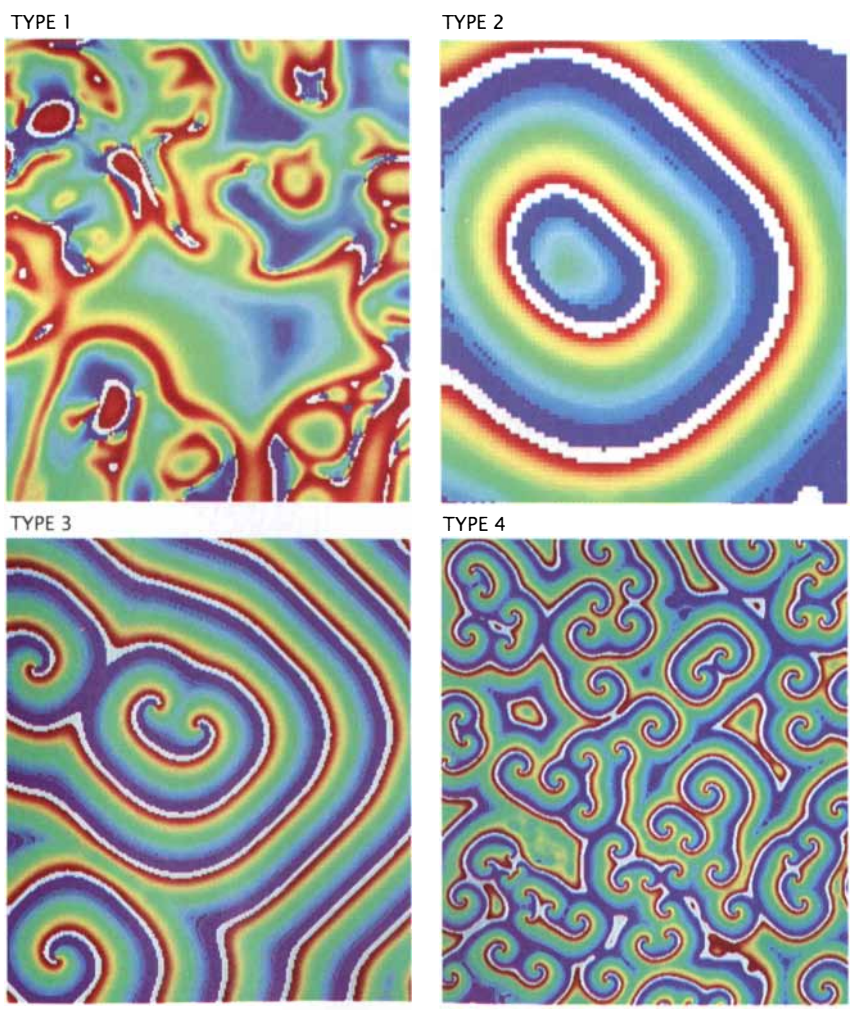
The next type of behavior they observed was labeled Type 2. It featured a generally regular series of infection "plateaus" roughly 30 cycles long, punctuated by the appearance of large numbers of healthy cells. (Sometimes nearly all 400 cells became healthy only to experience a new wave of infection.) As g was increased still further, Type 3 behavior appeared. It was heralded by the onset of a very regular alternation between saturation and virtual disappearance of infected cells every 20 cycles or so. Finally, Type 4 behavior emerged: within a few cycles of start-up the number of infected cells would fluctuate with some regularity about a saturation value of approximately 75 percent.

The four types of behavior appeared in order as g was progressively increased, but with some overlap: runs with transition values of g sometimes resulted in one type of behavior and sometimes in another type. In certain cases Gerhardt and Schuster even witnessed transitions between behaviors in a single run.

The four behaviors represent the appearance of specific types of wave patterns that are shown in the illustration below. In those color images the grid sizes vary from 100-by-100 cells to 500-by-500 cells. Waves associated with Type 1 behavior traveled only a short distance before dying out. Type 2 waves traveled outward in circular bands that varied greatly in width. Type 3 waves displayed the same circular shape but were more regular, in keeping with the regular ups and downs of infected cells displayed in its graph. Finally, Type 4 waves followed a spiral pattern that spread out from the center of the grid. As always, readers with computers are urged to repeat the experiment in some form. Waves of thought are sure to accompany the waves on one's screen.

Some of the wave patterns generated by the hodgepodge machine are similar to those displayed by a variety of chemical systems; certain ones in particular are dead ringers for the chemical waves found in the well-known Belousov-Zhabotinsky reaction. Compare, for example, the complex pattern of curlicues in the computer-generated image with the photograph of the Belousov-Zhabotinsky reaction in the illustration on page 107.

To what do we owe this similarity? Gerhardt and Schuster were not exactly surprised by it; they had deliberately designed the hodgepodge machine to mimic the features of a particular kind of "heterogeneous catalytic reaction" in which carbon monoxide and oxygen combine to form carbon dioxide while adsorbed at the surface of thousands of tiny palladium crystallites dispersed throughout a porous medium. Heat given off as the oxidation reaction proceeds changes the state of the catalyst. An abrupt phase transition by the crystallite liberates the carbon monoxide adsorbed at its



The hodgepodge machine produces distinctive wave patterns

surface; the catalyst then cools and the reaction begins anew.

The hodgepodge machine proved capable of mimicking not only this reaction but also the Belousov-Zhabotinsky reaction quite well. In the Belousov-Zhabotinsky reaction malonic acid is oxidized by potassium bromate in the presence of a catalyst such as cerium or iron. The grid cells of the hodgepodge machine in essence represent the catalyst particles, and the infection metaphor expresses the gradual saturation of the particles' surfaces.

But the analogy is not quite so simple; there are some subtleties here. For one thing, in the hodgepodge machine adjacent cells interact by exchanging infection, so to speak. How do the catalyst particles exchange reactivity? Gerhardt and Schuster reasoned that, at least in the case of the carbon monoxide oxidation, the participating catalyst units influence their neighbors by means of two basic mechanisms. A given unit could be made more reactive by the transfer of heat from a more active neighboring unit or by the diffusion of carbon monoxide from a less active neighbor.

The interaction between neighboring cells in the hodgepodge machine makes it possible for them to synchronize their activities. After a period of initial random disorganization (the hodgepodge phase), the patterns that appear reflect this synchronization. The same is presumably true of the actual chemical reactions as well. Does the hodgepodge machine thus explain the appearance of waves of excitation in the reactions it simulates?

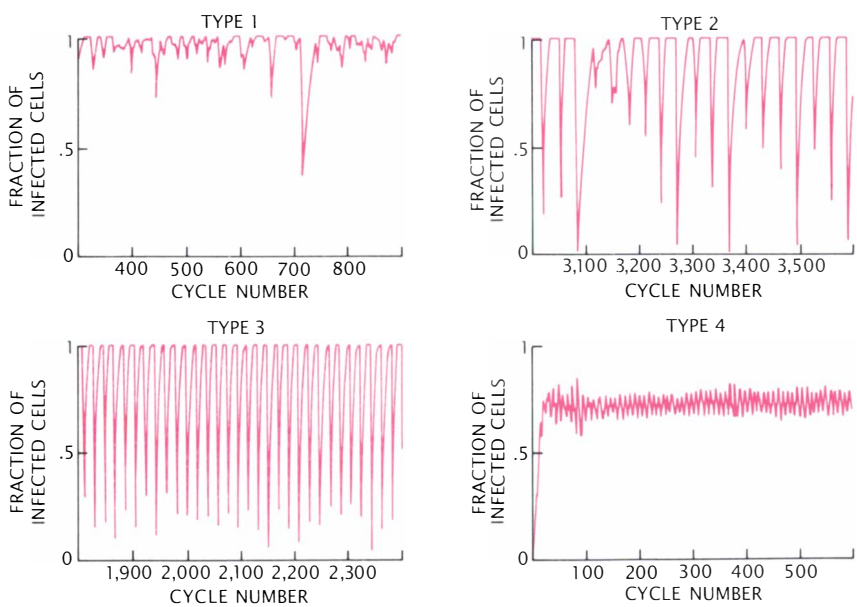
There will be those who are ready to exclaim "Of course!" and to point to the pictures as evidence. But then, there are people who see a cellular automaton in everything. In April *The Atlantic* carried an article about the cosmic ramblings of Edward Fredkin. A computer businessman and sometime academic, Fredkin supposes our universe to be composed of cells that tick from state to state like a vast cellular automaton. To be kind, the evidence for such an arrangement is not overwhelming. The hodgepodge machine is doubtless significant, but the attitude of its discoverers is more so. In spite of the fact that the hodgepodge machine simulates the Belousov-Zhabotinsky reaction remarkably well, Gerhardt and Schuster do not claim that chemistry is cellular. Instead they see their automaton as an approximation tool, the discrete version of a partial differential equation.

Originally inspired by the work of chemists Nils Jaeger and Peter Plath of the University of Bremen, Gerhardt and Schuster along with their mentor at Bielefeld, Andreas W. M. Dress, have enlisted the help of two chemists in studying the hodgepodge machine: S. C. Müller of the Max Planck Institute for Nutritional Physiology in Dortmund and John J. Tyson of the Virginia Polytechnic Institute and State University. The creators of the machine want to show that an array of chemical oscillators that interact locally according to certain simple rules will inevitably generate waves. Presumably there are only a small number of possible wave patterns, although they become far more complicated in three dimensions, according to Tyson. Because three-dimensional wavefronts are much harder to see in laboratory glassware, computer simulations may tell chemists what to look for. In science the trick is to use models well, not to be used by them.

Readers who would like to build their own hodgepodge machine have already received ample hints on how to proceed. One must declare an array of appropriate size and incorporate it into a grand loop that updates the array according to the three rules and then displays it for the edification of local hodgepodgers. Each element of the array must contain the state number for a particular cell. In computing the updated array, however, it is necessary to store the results temporarily in another array until the computation is complete. Then a simple double loop allows wholesale replacement of the original array by the updated one.

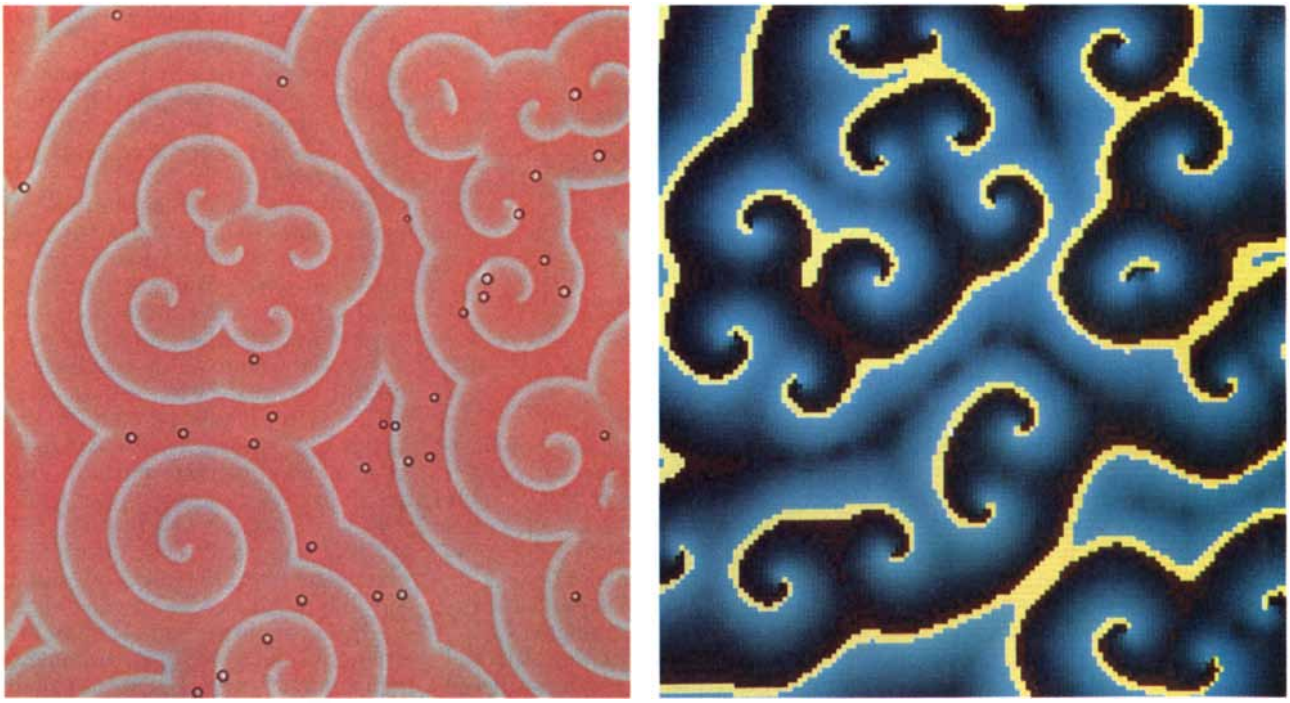
The updating is also carried out by a double loop. Two index variables, say *i* and *j*, count off the cells of the grid. For each cell given by the coordinates (*i*,*j*), the program (can we call it anything but HODGEPODGE?) decides by means of a pair of "if" statements whether the cell is healthy or infected. If it is healthy, the first formula is evaluated. If it is infected, the second formula is evaluated. In either case the states of the cells in its neighborhood must be checked. If the cell is neither healthy nor infected, it is obviously ill and will recover at the next cycle.

For reasons of space I am limited to this brief recipe. Readers who would like a more complete algorithmic description of the hodgepodge machine should write to me in care of this magazine. Please include a check or money order for $2 to cover postage (worldwide), copying and other costs.



The four behaviors of the hodgepodge machine

The Apraphulian excursion of April fooled few people. Those who nonetheless entered into the spirit of the account were challenged by the reconstruction of the Apraphulian analog multiplying machine: a device that multiplies two numbers entirely by means of ropes and pulleys. Some entered into the spirit of the enterprise so fully that they asserted they had firsthand knowledge of the ancient Apraphulian culture. The champion letter in this vein was sent in by Clive J. Grant of Chichester, N.H. A long document describes Grant's correspondence with a



Wave phenomena in a Belousov-Zhabotinsky chemical reaction **(left)** *and their hodgepodge counterparts* **(right)**

mysterious Dr. Ebur Grebdlog, renowned scholar of Apraphulian lore:

"After reading your 'Computer Recreations' column on the state of Apraphulian mathematics, I contacted Dr. Grebdlog to ask him if his work had ever covered. . .an Apraphulian Analog Multiplier. . . . Indeed, he replied, he had investigated that matter, and he sent along a copy of his work."

The "work" was beautifully written in Grebdlog's spidery hand, accompanied by technical drawings of pulleys and cams connected by bridges. Grebdlog notes that the drawings "appear to have guided the Apraphulians in constructing a device truly remarkable for the unstinting technological effort applied to its development but more remarkable for its total lack of utility."

The multiplier most often suggested by readers made use of a rod one end of which is attached to a fixed hinge. An input rope tied partway along the rod pulls it forward so that an output rope tied to the free end is also pulled in the same direction. Since the rod is in essence a lever with the fulcrum at its hinged end, the output rope moves a greater distance than the input rope. The problem with this design arises from a loss of proportionality: as the input rope is pulled farther, the rod follows a circular arc and the amplifying effect on the output rope eventually fades. Variations on this theme sometimes corrected for the rod's circular motion by means of guides or fancy systems of parallel jointed rods. All of this struck me as too complicated. Perhaps I should have explicitly forbidden the use of rods.

Robert Norton of Madison, Wis., used spiral pulleys to compute logarithms and antilogarithms. The input ropes *A* and *B* are unwound from two drums (which are not against the rules, since they are just wide pulleys). Each drum is attached to a spiral drum that winds up an output rope. The two outputs are then added in the way outlined at the end of the April column. The antilog of the sum is computed by winding the addition rope onto a spiral drum that is connected to a straight drum on which the final output rope is wound. A similar machine was "discovered" by Robert A. Eddius of New York City. The Apraphulians, he contends, used the shells of certain mollusks whose spiral shape enabled them to compute logarithms exactly! On the other hand, David A. Fox of Lima, Ohio, writes us that a similar culture inhabited a small island off the Marshall group known as Hardly Atoll. Here were found not only the same log-antilog devices but also a contraption rather like a yo-yo that was capable of squaring numbers. Readers might want to ponder whether Fox's assertion is possible.

Caxton C. Foster of East Orleans, Mass., is of the opinion that the Apraphulian civilization was destroyed by logical gain: the problem encountered by a computer in which the "1" output of each gate is not quite 1. To prevent such inaccuracies from creeping into the sacred computations, the high priests stationed an Apraphulian at each gate to pull a little harder on any output ropes lacking the necessary tautness. Thus absorbed, the people were unable to procure food and eventually starved to death.

The final word belongs to modern-day computer architect Michael Pagan of Mount Laurel, N.J. Concerned about the cultural gap between the analog branch and the digital branch of Apraphulian society, Pagan developed a marvelous analog-to-digital converter. A single rope carrying the analog signal enters the device and a number of ropes bearing the digital equivalent of the input number leave it. Such a machine may have been introduced on Apraphul, but the priests would certainly have banned the pagan device.

FURTHER READING

THE ARMCHAIR UNIVERSE. A. K. Dewdney. W. H. Freeman and Company, 1988.

DID THE UNIVERSE JUST HAPPEN? Robert Wright in *The Atlantic,* Vol. 261, No. 4, pages 29–44; April, 1988.

BOOKS

Psychic warfare, blood relatives, Clovis culture, wizards of the wastebasket

by Philip Morrison

ENHANCING HUMAN PERFORMANCE: ISSUES, THEORIES, AND TECHNIQUES, edited by Daniel Druckman and John A. Swets. Committee on Techniques for the Enhancement of Human Performance, Commission on Behavioral and Social Sciences and Education, National Research Council. National Academy Press ($32.50; paperbound, $22.50).

"Be all that you can be," those Army billboards urge. To turn many a young man or woman into a skilled technician or a soldier fit for battle is a major part of what the U.S. Army must do; but time is short, turnover is high, severe stress is certain and not all the learners are prepared to progress.

The Army's mission frees it from most social norms of schooling; what goes is what will work. It is no surprise that "some influential officers" are out to find new educational techniques, however extraordinary. Nothing can be rejected as implausible, not even the ability "to influence the operation of distant machines" by specific mental effort. Novel schemes often appear as well-promoted packages, supported by adherents who voice unstinted claims. Confidential reports abound that somewhere near the Urals such and such a new scheme has been tried; can we be so imprudent as to neglect it entirely? (No reference is made to an old disappointment of Reichsführer S.S. Heinrich Himmler, whose picked team of Aryan sensitives searched in vain for the Royal Navy with hand-held pendulums set swinging above a big map of the North Sea.)

This brief book is the wide-ranging and reflective report of an expert committee that was asked by the Army Research Institute to evaluate a set of extraordinary techniques that promise to better specific kinds of human performance. The 14 members brought long experience, and distinction earned both in and out of the psychology laboratory. Their specialties include attention, learning, memory, experiment design and even theater magic. They convened half a dozen times over two years, made site visits and commissioned a variety of special studies. Finally they issued this pithy account. In it basic psychological insights merge very readably with the critique of specific technique.

What works? Well, controlled and robust experimental results with 100 subjects confirm that it is possible to slow heat loss from the chilled hand. A few hours of biofeedback training suffice; by employing it the skin temperature can be elevated a few degrees. The benefits are improved dexterity and reduced pain. Heart rate under heavy exercise can be slowed 5 or 10 percent by several days of feedback training; that worked with monkeys too. So far no practical exploitation of these results has been achieved.

What does not work is learning during sleep. Subjects do learn material presented by sound well enough under such regimes. Yet once their sleep is monitored for alpha activity (the electroencephalographic sign of arousal), the improved recall is found to take place only with the sacrifice of real sleep. Maybe that would be worth the cumulative physical cost to the student. The application is not new: three weeks were saved in a 1916 course of Morse code for sailors. The panel concludes that the method "deserves a second look."

Some packaged schemes are more complex. Take SyberVision, a kind of mental practice designed to teach motor skills. The student repeatedly watches an hour-long video of a sports professional, shown making his skilled moves at varied speeds and viewing angles, as carefully chosen music plays. Certain imagined reenactments and changes of attention are enjoined on the student. There are other versions. It seems that such mental practice may produce half a standard deviation in performance gain in a subject when compared with controls. But the theories suggested by the proprietary developers of such systems—one offers an analogy to Fourier transforms—find little support. If some real physical practice is mixed in, the performance gains are much greater.

The promoters of Suggestopedia, a method devised by a Bulgarian psychotherapist, claim a 25-fold acceleration in language learning. It is a many-sided scheme: comfortable chairs instead of desks in rows, a few minutes of relaxation and recall of pleasant learning experiences in the past, careful, dramatized presentation of the material. A passive review conducted in time to baroque music is followed by a final practice and quiz session. It sounds worthwhile. But someone compared seven instructors who taught using both Suggestopedia and the "silent way," a method characterized by little verbalization or repetition and a tense environment. Both approaches worked about equally well; the only significant variable was the instructor. Half a dozen other schemes are reported on; one or two hold some promise.

A closing chapter seriously treats the paranormal. The committee made a determinedly open-minded effort to examine claims that might lead to psychic intelligence gathering or even psychic warfare. Members visited three laboratories where psychic effects are now under study and carefully reviewed the ample literature. First they report on the evaluation of extrasensory perception. Adepts were tested by matching their written descriptions to real places the judges could visit. "In summary, after approximately 15 years. . .only one possibly successful experiment. . .provides only marginal evidence for the existence of ESP." The enthusiasm of proponents rests on the exaggerated claims for early experiments and on "illusory" correspondences found in parts of the descriptions.

Second, the committee examined mental efforts to induce electronic random-number generators to issue nonrandom sequences. Over the years most such experiments have been done under conditions that do not meet "the minimal criteria of scientific acceptability." Just one of some 200 experiments, reported in 1980, was "singularly well controlled." It yielded results of marginal significance; there was a probability of one in 30 that the observed results could have been gained by chance. A long series done more recently at Princeton has few weaknesses: the hit rate was 50.05

In case you missed Audio Update™ back issues, these important and timely programs are still available:

March/April 1987:

COCAINE, featuring Jeffrey Isner, M.D., Tufts University School of Medicine and New England Medical Center.

AIDS, featuring Robert C. Gallo, M.D., National Cancer Institute; Martin S. Hirsch, M.D., Harvard Medical School and Massachusetts General Hospital.

TUMOR IMMUNOLOGY, featuring Herbert F. Oettgen, M.D., Memorial-Sloan Kettering Cancer Center; Steven A. Rosenberg, M.D., National Cancer Institute; Stephen F. Lowry, M.D., New York Hospital–Cornell Medical Center.

May/June 1987:

HBLV: THE NEW HERPESVIRUS, featuring Robert C. Gallo, M.D., National Cancer Institute.

ALCOHOLISM, featuring Enoch Gordis, M.D., National Institute on Alcohol Abuse and Alcoholism.

ARRHYTHMIAS: TECHNOLOGIC ADVANCES, featuring Jeremy Ruskin, M.D., Harvard Medical School and Massachusetts General Hospital.

GENETIC DISEASE, featuring Helen Donis-Keller, Ph.D., Collaborative Research, Inc.; Louis D. Kunkel, Ph.D., The Children's Hospital, Boston.

July/August 1987:

PROTOZOAN INFECTIONS, featuring Peter F. Weller, M.D., Harvard Medical School and Beth Israel Hospital, Boston.

MALARIA VACCINE, featuring Ruth S. Nussenzweig, M.D., Ph.D., and Victor Nussenzweig, M.D., Ph.D., New York University Medical Center.

THE IMMUNE SYSTEM, featuring Philip Leder, M.D., Harvard Medical School and Howard Hughes Medical Institute; Hugh McDevitt, M.D., Stanford University School of Medicine.

September/October 1987:

CHOLESTEROL AND HEART DISEASE, featuring Claude Lenfant, M.D., National Heart, Lung, and Blood Institute.

CANCER THERAPY AND MULTIDRUG RESISTANCE, featuring Michael Gottesman, M.D., National Cancer Institute.

ANTIMICROBIAL CHEMOTHERAPY, featuring Harvey B. Simon, M.D., Harvard Medical School and Massachusetts General Hospital.

RETROVIRUSES AND HUMAN DISEASE, featuring Myron Essex, D.V.M., Ph.D., Harvard University School of Public Health.

ALLERGIES, featuring Lawrence M. Lichtenstein, M.D., Ph.D., Johns Hopkins University.

MENINGITIS, featuring James J. Rahal, M.D., New York Infirmary–Beekman Hospital and New York University.

CANCER EPIDEMIOLOGY, featuring Robert N. Hoover, M.D., Sc.D., National Cancer Institute.

CARDIOMYOPATHIES, featuring Roman W. DeSanctis, M.D., Massachusetts General Hospital and Harvard Medical School.

January/February 1988:

ANXIETY DISORDERS, featuring Gerald L. Klerman, M.D., Cornell University Medical College and Payne Whitney Clinic.

HEMOSTASIS, featuring Lawrence Leung, M.D., Stanford University School of Medicine.

ACUTE MYOCARDIAL INFARCTION, featuring Thomas W. Smith, M.D., Brigham and Women's Hospital and Harvard Medical School.

LYME DISEASE, featuring Allen C. Steere, M.D., New England Medical Center and Tufts University School of Medicine.

March/April 1988:

ALZHEIMER'S DISEASE, featuring Peter Davies, Ph.D., Albert Einstein College of Medicine.

OSTEOPOROSIS, featuring Robert Neer, M.D., Massachusetts General Hospital and Harvard Medical School.

RESPIRATORY FAILURE, featuring Roger C. Bone, M.D., Rush-Presbyterian-St. Luke's Medical Center and Rush Medical College.

PEPTIC ULCER DISEASE, featuring Charles T. Richardson, M.D., Dallas Veterans Administration Medical Center and University of Texas Southwestern Medical School.

May/June 1988:

SPORTS MEDICINE: EXERCISE AND HEALTH, featuring Harvey B. Simon, M.D., Massachusetts General Hospital and Harvard Medical School.

URINARY TRACT INFECTIONS, featuring Robert H. Rubin, M.D., Massachusetts General Hospital and Harvard Medical School.

ASTHMA, featuring Thomas A. Raffin, M.D., Stanford University Medical Center.

ANTIFUNGAL THERAPY, featuring Richard D. Diamond, M.D., University Hospital and Boston University School of Medicine.

Each tape costs only US$14.95 (US$15 outside the U.S. and Canada). Just indicate your choices on the coupon below.

Send for your Audio Update™ tape series today—and start listening to the voices of medicine.

SCIENTIFIC AMERICAN MEDICINE AUDIO UPDATE™

A 8

415 Madison Avenue, New York, NY 10017

I'm ordering:

- ☐ March/April 1987
- ☐ May/June 1987
- ☐ July/August 1987
- ☐ September/October 1987
- ☐ November/December 1987
- ☐ January/February 1988
- ☐ March/April 1988
- ☐ May/June 1988
- ☐ July/August 1988

at US$14.95* each (**Note: single tape orders require prepayment by check**).

☐ **Subscription:** start my yearly subscription with the current program. I'll receive 6 bimonthly tapes for US$59.95* (a discount from the single-issue price).

☐ Check enclosed* ☐ VISA ☐ MasterCard ☐ Bill me

Exp. date _____ Account No. _____

Name _____

Address _____

_____

_____ Zip Code _____

Specialty _____

*Prices apply to U.S. and Canada. Please add sales tax for DC, IL, MA, MI, or NY. Allow two weeks for your order to be processed.

Prices outside U.S. and Canada: Single tape—US$15; Subscription—US$75. Please make payment in U.S. dollars.

percent, but the results are in some doubt, since statistical significance rests on data for one person only, who was conversant with the experimental routine and familiar with the equipment. Tampering cannot be excluded. Even in the one best experiment the innocent presence of the subject might have biased the output very slightly by some physical effect. In the control experiments the random-number generator was in an empty room. The miracle of mind over matter has faded into at most a faint whisper of statistics.

A third visit witnessed the sympathetic polygraphic response to a distant subject's emotional state shown by a suspension of cells that had been taken from the subject's mouth. There was little signal in the noise.

Mental spoon bending began as a trick of the conjurer. Now there are devotees of psychokinesis who organize "PK parties" that achieve "a state of emotional chaos." Everyone shouts "Bend!" and people avoid looking at what their hands are doing. These enthusiasts are not few. "Over and over again we have been told by participants that they know that metal became paranormally deformed." Of course not one documented case was presented. Such self-persuasion is vivid; one rowdy memory of distortion outweighs any number of controlled failures. Just there lies a sufficient reason for doubt. Among the most difficult lessons in science is how not to deceive yourself. This patient and judicious overview offers genuine help. A useful glossary and references are provided.

ATLAS OF BLOOD CELLS: FUNCTION AND PATHOLOGY, edited by D. Zucker-Franklin, M. F. Greaves, C. E. Grossi and A. M. Marmont. Second edition. Edi. Ermes, Milano, and Lea & Febiger, Philadelphia ($225).

The two large colorful volumes are the work of a team of about 20 collaborators from Italy, the U.S. and the U.K. (Dr. Greaves of London). The investigators have prepared a visually centered study of what is known about the cells that live and die in red blood and clear lymph. Here are *T* cells from the thymus and *B* cells from the bone marrow, what they do, how they appear and how they sometimes fail. The up-to-date work is aimed at clinicians, with the explicit intention of transmitting the burden of the mushrooming literature by carefully made diagrams and tables, along with plentiful photomicrographs of cells and their ultrastucture. Whereas the authors and editors sought only scientific excellence, Dr. Italo Grandi, who guided the Italian printers and photolithographers, has produced a work of art.

Nearly 800 big pages present more than a general reader—or a reviewer—is able to absorb. Yet the strongly visual presentation of the new partnership between modern immunobiology and the more familiar microscopic examination of stained tissue and blood samples is striking even to a bystander. That is all the more clear as we catch sight of how matters stand over a wide spectrum of pressing problems and brand-new approaches. Genetic errors, malignancies, infections of the blood-forming organs, AIDS, bone-marrow transplants, commercial monoclonal antibodies, even automatic cell-sorting—all and more are presented.

Blood cells come in great variety. They share common ancestors: the stem cells of marrow or lymph. Those ancestral cells are only inferred by techniques of tissue culture that disclose differentiation. After a step or two of division, eight recognizable classes of differentiated descendants emerge. Empirical dye staining and the power of long-trained human pattern recognition still guide everyday diagnosis in this domain, but at last they have real allies.

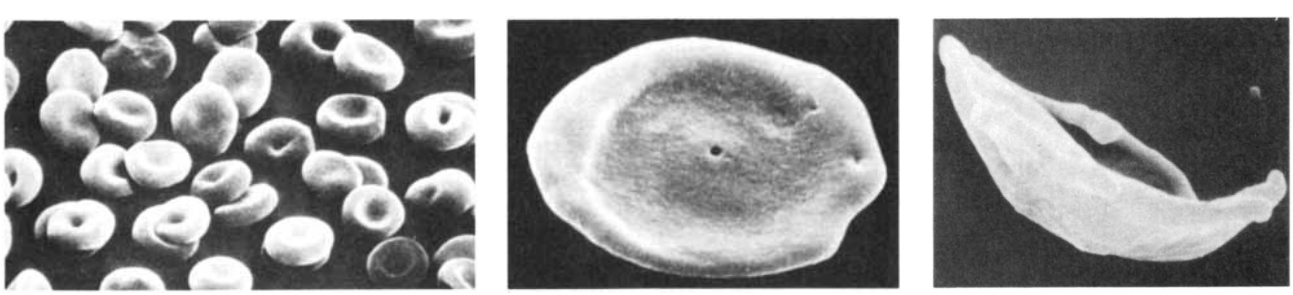
The surfaces of the target cells bear molecular receptors, and specific antibody markers can be grown using clones of the right genetically engineered microbial cell. Those antibody molecules can be labeled for viewing by eye and film. Unknown cell components were once chemically stained by a reactive dyestuff. Now a cell-membrane protein picks up the specific partner protein that fits, and with it an incorporated fluorescent dye.

The marked white cells, leukocytes, show up superbly under the microscope; some glow green, some orange, seen in ultraviolet like a star cluster against a dark ground. Forty-five such marked antigens can be bought commercially nowadays. A table lists them all, along with the expected distribution of their receptors among the various blood cells.

Volume 1 treats red cells at length; the importance of these enucleate, short-lived specialists is evident. Their forms are diagnostic for a variety of disorders, anemias and plethoras too. Iron-deficient specimens are ragged and unstainable. A single error in the long chain of amino acid monomers that link into protein for hemoglobin can collapse the little biconcave disks to crescents and sickles. Here is a marvelous shot of a tiny crystal, secreted within the infrequent white blood cells called eosinophils. Those cells draw on the crystalline store of a toxic protein to attack parasitic protozoa encountered in the blood. We see such a cell smearing its binding poison onto a schistosome. The volume goes on to treat what is known and visible about the other less evident classes of white blood cells.

The lymphocytes occupy most of volume 2. The older ways of detection were not less exciting, but they were less sure; here is a set of fine photographs of rosettes, clusters of foreign red-cell disks seen held to lymphocytes having the right receptors. Chapter after chapter emphasizes the variety of these white cells. Their pathologies are often diagnosed by an excess of a single surface marker and by subtle telltale patterns of the stained tissues of the lymph nodes the cells populate.

The human immunovirus is not easy



***Normal red cells* (left), *a beta-thalassemic cell* (middle) *and a deoxygenated sickle cell* (right)**

to detect visually, even though the leukocytes are its chief target. One of the deadliest of the opportunistic invaders in AIDS, the parasite *Pneumocystis carinii,* is visible in 90 percent of autopsies on AIDS victims, even though the right antibody is abundant. The organism is present in healthy lungs as well; we do not know what frees the parasite from host control.

A remarkable chapter reviews the chromosomal abnormalities seen in the leukemias. More than any other tumor, these cells of blood and marrow show frequent gain and loss of whole chromosomes. They acquire strange bands of nucleoprotein. They develop specific errors of rearrangement during descent. We are close to recognizing specific oncogenes, the misprints or ambiguous words in the long cellular instructions that are somehow the first of many steps to proliferative crisis. This hopeful naval campaign among the floating cells, where our visual and chemical reconnaissance is now so good, might lead to eventual victory in the great war against all the cancers.

PREHISTORY OF THE AMERICAS, by Stuart J. Fiedel. Cambridge University Press ($49.50; paperbound, $14.95).

This concise, clear, rather austere paperbound bargain tells the history of our species in the New World. It is even more of a find for the serious general reader than for the undergraduates for whom it was written. Trained as an Old World archaeologist, and therefore something of the knowing outsider, Dr. Fiedel was emboldened by what he read to attempt the "rather grandiose" task. His success can be easily and enjoyably shared. What he does so well is offer an explicit account of key issues. For each he gathers and compares arguments and then supports what he says by presenting the cogent details of sites, artifacts and cultural complexes. The style contrasts with one that places some important new theory at the center and then seeks to support it. No one will think that the last word has been said, here or elsewhere, on most of the great questions addressed.

A chapter on the history of American archaeology opens the book. Next comes a portrait of human beings in the Paleolithic. The setting moves half a world northeastward from the Rift Valley to Ushki on the Kamchatka Peninsula. There, near the gateway to the Americas, a site yielded stone tools and underground huts that date from 12,000 B.C.

Since Fray José de Acosta suggested the idea in 1590, acceptance of the proposition that the first Americans came from subarctic Asia has grown to virtual unanimity. They brought Mongoloid genes, say for black hair and high cheekbones, but the founding sample was not a random one: among American populations the frequency of blood groups is not the same as it is for Asians, except among the Eskimo. There are a few enigmatic finds of human remains with possible old dates, but not even one that is of truly high antiquity, like those of Africa, or those of our ancestors who newly had fire, say in the caves of Choukoutien. This is a New World; all Americans were modern *sapiens* from the start.

The Bering route was dry land on and off until it flooded over to stay, some 10,000 or 15,000 years ago. An ice-free corridor to the south dates from about that time too. The migration could have begun at most 40,000 years ago, at the time our own subspecies rose to dominance. The peopling of the Americas could hardly have been delayed beyond that final submergence of hunters' trails to Alaska.

Just when our species first made the crossing is in doubt. There are chopper tools in plenty; they hint at great age, but most of them are surface finds, hard to date. By about 9500 B.C. there are many sites of a culture called Clovis, marked by its beautiful, expertly fluted large spearpoints for big game. Human remains and points that derive from that same style have been found all the way south to land's end in far Tierra del Fuego. These artifacts date from a little later than 9000 B.C.

"The most convincing evidence of pre-Clovis occupation" is that found at the Meadowcroft rock-shelter in western Pennsylvania, but it is not beyond reasonable doubt. The radiocarbon dates fall in sequence back to 17,000 B.C. The tools—no big spearpoints at all—might be post-Clovis or pre-Clovis in style.

Why should the Clovis lifeway have spread so swiftly? In general it resembles what we know from old Siberia—mobile bands equipped with big-game hunting gear—but it has no close Asian antecedents. The fluted point must be an American invention, probably made early during the migration southward. There is a well-known hypothesis for the rapid spread across the whole of the New World of those first capable hunters. It presumes that the extent and speed of the occupation was the consequence of abundant big game, still unwary of man. The idea is plausible but not yet proved.

Proponents of the theory cite the extinction of large species as evidence. True, the mammoth and the giant sloth were taken widely, and they vanished, but not the bison, moose or caribou. Nor did the hunters of Africa slaughter their big prey to extinction. Yet what is the alternative? Neither disease nor climate change seems likely to end most large mammals from the Tropics to the snow.

There is a wonderful find, called Monte Verde, in Chile, where mastodon hunters dwelt in wood huts; they seem to have used bolas and not spears. We need to learn more; the carbon dates there are a millennium or two earlier than Clovis, although the culture hints at a later time. If Monte Verde and other South American sites really antedate Clovis, why is nothing like them seen among a hundred finds farther north?

We are coming to know more of the hunters' ways. The finds can be differentiated into single kill sites where a few mammoths were speared in ponds or bog; larger, less transient, seasonal hunting camps, and base camps whose living floors were occupied by an entire band. At a kill site in Colorado of about 8000 B.C., bison were slaughtered by the herd, perhaps by being driven between banks of snow. A posthole was found there, and near it "an antler flute, a miniature point" and other objects associated with a shaman. A shaman has been seen to play just such a propitiating role at the seasonal hunt among Plains Indians: the potent strategy survived the test of time for 10,000 years.

How and why did the early Americans come to settle on the land and build villages? How did they pass over to life by farming? How did the chiefdoms arise, groups coherent enough to build the great mounds and pyramids that shoulder up from Ohio to Panama? How did those groups grow to the level of warring states, consisting of classes of subjects who ranged from priest-kings to artisans of luxuries? Can we understand the city-state of the valley of Mexico and the diffuse empire of control established so swiftly by the Inca, under Viracocha and his son, ominously named Cataclysm? The comparisons at every step are fruitful. Finally, is the peopling of the Americas the work of leaders from overseas, some Chinese genius perhaps, or is it an independent case of human behavior in the large?

All these issues are laid out, and tentative decisions are reached by reason: settlement preceded farming, gathering and growing are a continu-

um, chiefs arise in more than one way, states appeared without clear overseas influence. The swift terminal conquest by the Europeans owed relatively little to the advantage of horse and steel sword against the noble warrior clans with their cotton armor and obsidian blade. The victory belonged more to the infectious microorganisms that had evolved in the Old World during the long history of the domestication of farm animals. Biology turned out to be destiny, made so both by evolved antibodies and by early social innovation.

There are many line drawings, from a radiocarbon calibration to fitted monoliths at Cuzco, and lots of maps. It is the meaning of the past that is the center, even if all is not yet in focus; the beauty remains implicit in the grandeur of the puzzles.

FIRST LIGHT: THE SEARCH FOR THE EDGE OF THE UNIVERSE, by Richard Preston. The Atlantic Monthly Press, distributed by Little, Brown and Company (Inc.) ($18.95).

"The world of astronomy held three types of people: observers comfortable with telescopes; theorists comfortable with pencil and paper; and instrument builders comfortable with wires." Indeed, the reader will meet young members of a fourth moiety, those at home with the computer screens and the printouts that follow from a thousand intricate pages of digital code. Richard Preston, himself a subtle craftsman of the word, has written an admiring report of his year's ethnographic visit to the astronomers at Caltech and Palomar. His story turns on the shared craftsmanship that binds that purposive, self-conscious and diverse community.

The 200-inch Hale telescope within its Art Deco dome, through which first light passed in 1947, is the set for most of the action. The Big Eye is a heavy and looming presence, "masterwork of Depression engineering. . .colossal, welded, gray, aloof, . . . agile, . . .the climax of dreadnought design." We grow at home there amidst its blemishes and its wonders. We see the few tablespoons of Flying Horse telescope oil that bleed each night from the huge bearing to the floor. We join Preston five stories up at the cold mouth of the open tube where he admires a perfect illusion, the floating image of a sheet of stars hanging before his eyes.

That was no working session, but a guest's tour. Normally the sheltered observers, accompanied by night music on the loudspeakers, watch and comment on what the video screen displays as the motion of the earth tracks a hairline arc around the sky. Faint galaxies parade past in their variety: spindles, spirals and glowing balls. That television-amplified view is typically so deep that only an occasional object passes by that anyone has caught sight of before.

It is James Gunn's new digital camera with its four imaging chips—they call it the four-shooter—that looks up at a distant secondary mirror through a hole in the big primary to form the video spectacular. Of course, the device conscientiously stores on tape whatever it sees, picture element by picture element, at a couple of megabytes per minute. Every night during the dark of the moon half a dozen such small parties of explorers stare deep into look-back time from big telescopes on mountainsides around the world. Yet the plainest message from this correspondent with one of the boldest parties is not of frontier independence but rather of belonging.

It is the state of the art that counts. The subtly etched chips of crystal silicon that store and sort the faint signals electron by electron offer an example. They are reject chips that Gunn cadged from the fabricators. The growth of chip mastery was prerequisite; it was reconnaissance-satellite technology for which the makers had first developed their skills under defense contracts. But the Hale camera looks up, not down.

The four-shooter—from a distance it "looks no bigger than a rivet fastened at the bottom of the Hale"—is a 1,500-pound tubeful of quartz mirrors, gold connectors and stainless plumbing, of cheap foam insulation and discount motors. Indeed, it was designed and built in a basement shop by Gunn with a group of engineers and technicians known at Caltech as the Wizards of the Wastebasket. They are true *bricoleurs,* "experts in the arcana of trash." The tiny Swiss motors they find at lunch hour as they sort through the surplus bins of C&H Sales in Pasadena work better than those their aerospace suppliers offer for slow delivery at 20 times the price.

The 200-inch is a tie to an earlier age of robust technology. The newer big telescopes are steered digitally and adjusted electronically. Mirror, dome and mount are light in weight. The devices seem less like the *Queen Mary,* more like spacecraft. The Hale relies on unique mechanical computers. For instance, 36 lever mechanisms, each consisting of a few hundred moving parts, act as mirror supports; within them balanced weights passively shift to press on and trim the heavy glass mirror as it moves. They were built in 1948 and since then have been left almost untouched, rather in awe. They were oiled once; the result, judged by the images of stars, was not happy. Bruce Rule, their designer, is not worried. "'We didn't give ninety-day guarantees,' he said. 'We built for life.'"

Juan Carrasco is the philosophical senior night operator at the telescope, nine years of devoted apprenticeship behind him. Once he followed a different craft. As a young barber back in Texas, he learned to shave even the winos with a sure hand.

It is he who taps just the right small d.c. motor when the mirror cover will not open and knows when to change the tubes in the analog computer that normally causes the massive dome to follow the telescope. Digital computers are beginning to replace the old originals. The engineers, however, are not yet willing to give over control of the fast slewing motions to any computer; someday it might fail and crash the irreplaceable instrument. Carrasco, a prudent man, would never do that; he used to carry his infant daughters around on pillows. The musing, disputatious astronomers are certainly to be kept away from the telescope controls.

Astronomy is not all remote galaxies. Gene and Carolyn Shoemaker entered by way of the rocks, astronomy from geology, in pursuit of the relics of impact on earth and moon. Lately they have been using the 18-inch Schmidt—the Little Eye—to patrol for nearby asteroids and comets. The wide-field Little Eye "bulldozes the sky" for moving orbital neighbors, while the 200-inch "drills holes into lookback time." Carolyn searches big film after film under her stereo magnifiers with keen and tireless eyes. The wary discoverer of many a comet and asteroid merits a final word. She finds galaxies merely confusing. "The fainties can look like comets. I get so excited. Then I find out it's only a galaxy."

The book is in form and style rather like a novel. It chronicles the careers of a dozen engineers and astronomers associated with Palomar. Some of them are men long gone, recalled half as folklore: Walter Baade, George Hale, Rudolph Minkowski, Bernhard Schmidt, Fritz Zwicky. Some are the men and women the author traveled and visited with in 1986. These are crisp, evocative characterizations, both of personality and its origins and of the scientific issues they spoke about all night long.

TOYOTA CRESSIDA

ELEGANCE: NO PROBLEM.
Sumptuously appointed—automatic climate control, power windows and door locks. Theft Deterrent System, rich velour comfort or optional leather.

FREE CRESSIDA BROCHURE
No trouble at all! Just send your name and address to: Toyota Motor Sales, U.S.A., Inc., Cressida Brochure Offer, 750 W. Victoria Street, Compton, CA 90220

THE MOST TROUBLE-FREE NEW CAR SOLD.

Toyota Cressida is the most trouble-free new car sold in the U.S.* a desirable trait in any car. But compare the 1988 Cressida to luxury cars costing thousands more, and it becomes more desirable than ever. Its sophisticated 6-cylinder twin cam EFI engine provides quick response, a deep reserve of power and ceaseless cruising ability. Optional Toyota Electronic Modulated Suspension (TEMS) turns concrete to velvet. Exquisitely styled aerodynamics. And a lavish cabin with superior comfort for five. 1988 Toyota Cressida. Afford yourself the luxury of the most trouble-free car in America. Pure quality.

Get More From Life... Buckle Up!

PERFORMANCE: RELY ON IT.
From the leather-wrapped wheel to the Technics** AM stereo/FM stereo electronic tuning radio with cassette and acoustic flavor/tone control equalizer, luxury performs.

*Based on problems encountered in first 90 days of ownership—J.D. Power and Associates 1987 New Car Initial Quality Survey.
**Technics is a trademark of the Matsushita Electric Industrial Co., Ltd.
© 1987 Toyota Motor Sales, U.S.A., Inc.

TOYOTA QUALITY
WHO COULD ASK FOR ANYTHING MORE!

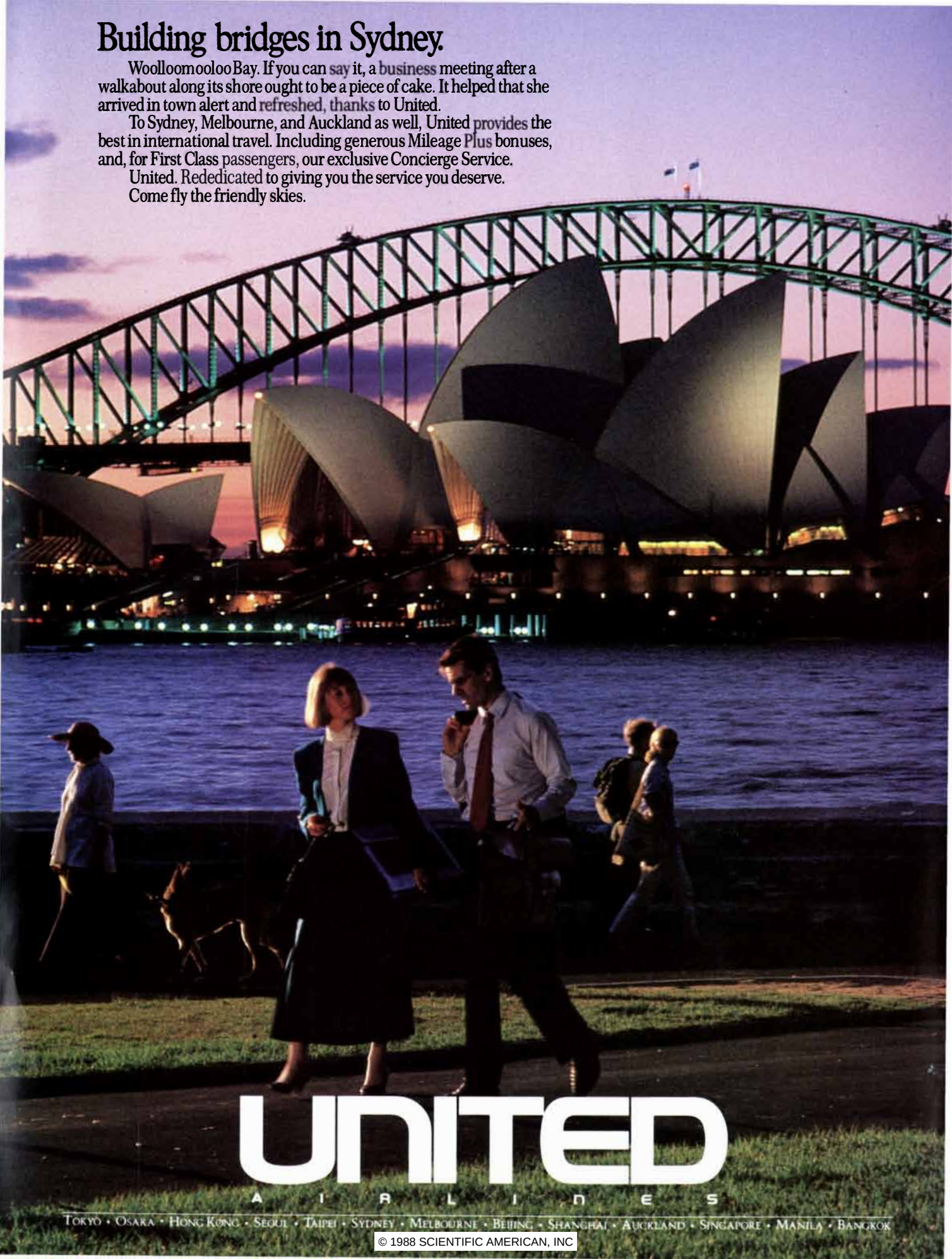
Building bridges in Sydney.

Woolloomooloo Bay. If you can say it, a business meeting after a walkabout along its shore ought to be a piece of cake. It helped that she arrived in town alert and refreshed, thanks to United.

To Sydney, Melbourne, and Auckland as well, United provides the best in international travel. Including generous Mileage Plus bonuses, and, for First Class passengers, our exclusive Concierge Service.

United. Rededicated to giving you the service you deserve.

Come fly the friendly skies.



UNITED

AIRLINES

TOKYO • OSAKA • HONG KONG • SEOUL • TAIPEI • SYDNEY • MELBOURNE • BEIJING • SHANGHAI • AUCKLAND • SINGAPORE • MANILA • BANGKOK