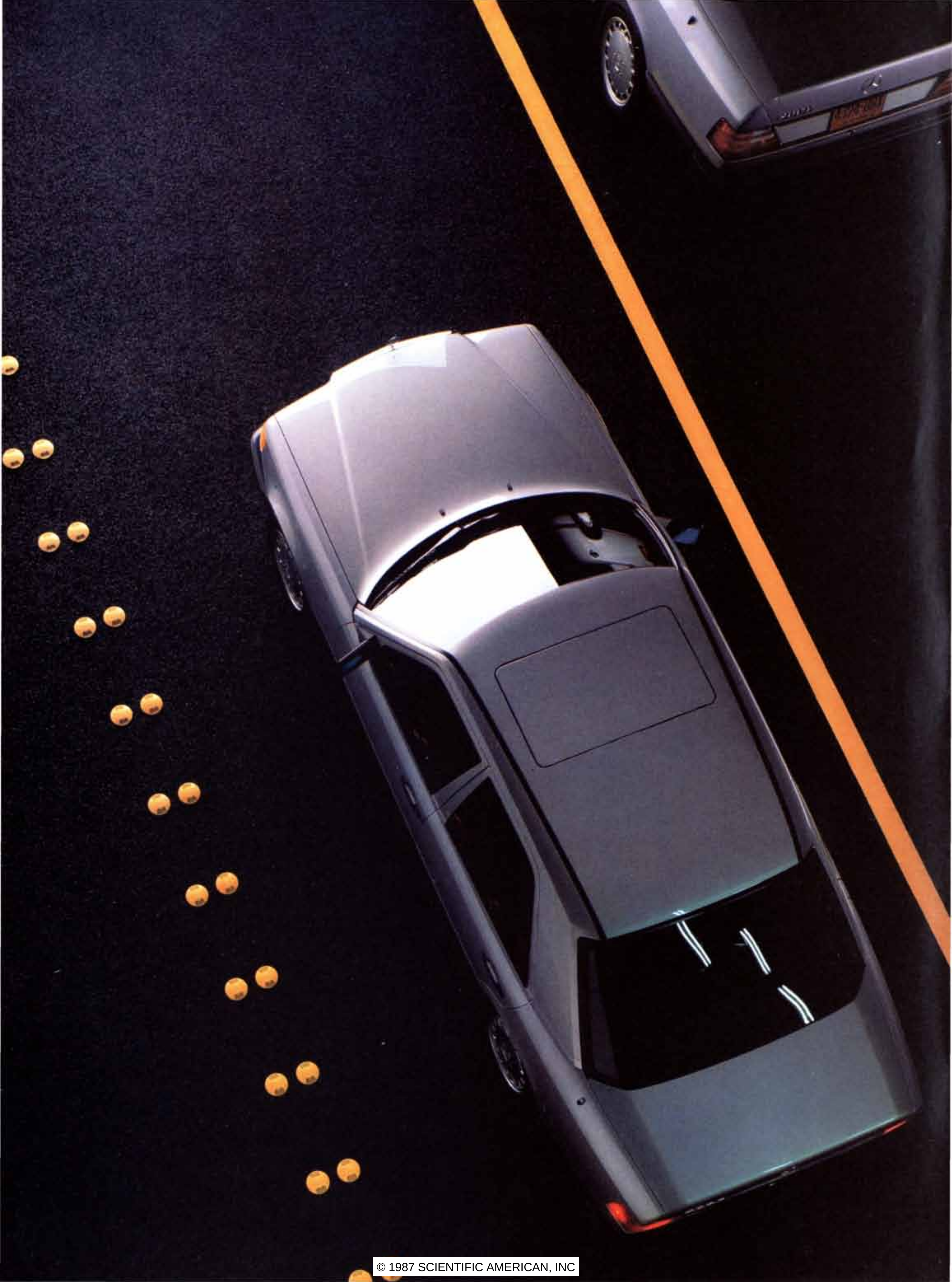
SCIENTIFIC AMERICAN

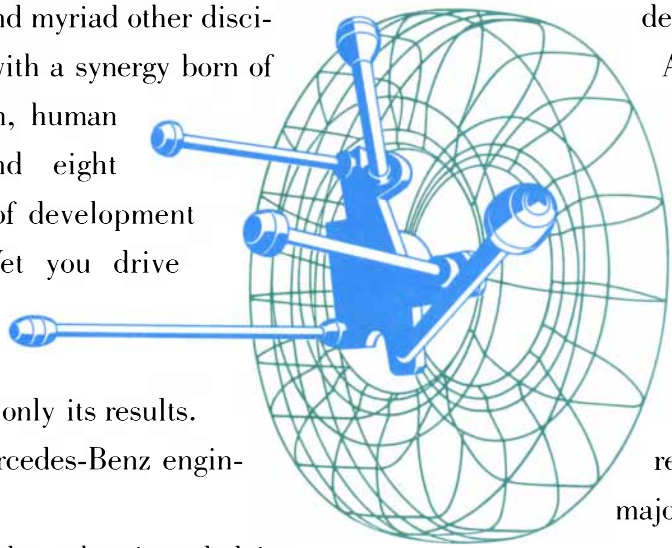
DYSLEXIA

$2.50

March 1987



THE MERCEDES-BENZ 300 CLASS: SO TECHNOLOGICALLY ADVANCED, YOU BARELY NOTICE THE TECHNOLOGY.



Microcircuits silently work their millisecond miracles. Hydraulics, pneumatics, kinematics, aerodynamics and myriad other disciplines interact with a synergy born of computer design, human inspiration, and eight arduous years of development and testing. Yet you drive oblivious to all this technology, savoring only its results. Precisely as Mercedes-Benz engineers intended.

Technology that intruded into your driving life could never have found its way into the 300 Class. Because it would then run counter to the sole justification for technology—enhancing the car's ability to transport its occupants with maximum efficiency, pleasure and safety.

Thus, you might never notice the five individual steel links that guide the movements of each rear wheel, comprising the most sophisticated steel suspension extant, as illustrated above. But you do notice their effect: handling prowess sufficient to "challenge many high-priced sports and GT cars," according to one automotive journal. A ride on rough surfaces that is "nothing short of magical," in the words of another.

Onboard electronics serve higher purposes than orchestrating dashboard pyrotechnics: continuously recalibrating the gasoline fuel-injection system for changing conditions and driver demands. Governing the ABS Anti-lock Braking System to help thwart wheel lockup during sudden hard stops. Monitoring the Mercedes-Benz Supplemental Restraint System (SRS), ready to deploy its driver's-side air bag and front seat belt emergency tensioning retractors within an eyeblink of a major frontal impact.

Nothing that you see or hear or feel seems overtly "technological." And yet that aerodynamically immaculate body all but banishes wind noise from the cabin. Your seat still feels comfortable after hours of driving. You find that every control falls logically to hand. And that your outward vision is reassuringly clear even in a rainstorm.

Advanced technology is at work all around you, but you do not think of it as technology. You think of it as the rewards of driving a 300 Class Mercedes-Benz.

Engineered like no other car in the world


A Grand Setting.
Grand Marnier Liqueur
750 ml. (25.4 fl.oz) - 80° Proof
For gift delivery anywhere call 1-800-CHEER-UP (except where prohibited by law).
Product of France. Made with fine cognac brandy 80 proof. ©1986 Carillon Importers, Ltd., Teaneck, N.J.

Established 1845 **SCIENTIFIC AMERICAN** March 1987 Volume 256 Number 3

ARTICLES

DEPARTMENTS

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 RATE: $24 PER YEAR IN THE U.S. AND ITS POSSESSIONS; $33 PER YEAR IN ALL OTHER COUNTRIES. POSTMASTER: SEND ADDRESS CHANGES TO SCIENTIFIC AMERICAN, 415 MADISON AVENUE, NEW YORK, N.Y. 10017.

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, UT 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, MA 01970. Specify CCC Reference Number ISSN 0036-8733/86. $1.25 + 0.00.

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

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

Subscription correspondence should be addressed to Subscription Manager, SCIENTIFIC AMERICAN, P.O. Box 5969, New York, NY 10017. The date of the last issue on your subscription is shown in the upper 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 ________________

THE COVER

The photograph on the cover depicts mirror writing. Commonly taken to be a symptom of dyslexia, mirror writing is often seen in young children who go on to develop normal reading and writing ability (see "Dyslexia," by Frank R. Vellutino, page 34). The drawing and its full legend ("My blanket I love the best") were actually produced by a four-year-old girl, whose progress from mirror writing toward normal writing is traced on page 35.

THE ILLUSTRATIONS

Cover photograph by Quesada/Burke

Page	Source
16–22	Andrew Christie
35–36	Frank R. Vellutino
37	Macdonald Critchley and Eileen A. Critchley
38–41	Johnny Johnson
43	Arthur J. Olson, Research Institute of Scripps Clinic
44	George V. Kelvin (*top*); Tzyy-Wen Jeng, University of Arizona (*bottom*)
45	George V. Kelvin (*top*); James M. Hogle, Marie Chow and David J. Filman, Research Institute of Scripps Clinic (*bottom*)
46	David J. Filman
47–48	George V. Kelvin
49	David J. Filman
51	William D. Phillips
52–55	Hank Iken
56	William D. Phillips
78–80	Grant Heilman Photography
81	J. C. Allen & Son, Inc.
82–86	Joan Starwood
89	Eung G. Paek and Ken Hsu, California Institute of Technology (*top*); George Retseck (*bottom*)
90	George Retseck
91	John Hong, California Institute of Technology (*top*); Thirumalai Venkatesan, Bell Communication Research, Inc. (*bottom left*); George Retseck (*bottom right*)
92	George Retseck
93	Patricia J. Wynne
94	Eung G. Paek
96	Enrico Bonatti, Lamont-Doherty Geological Observatory
98–100	Andrew Christie
101	Andrew Christie (*left*); Enrico Bonatti (*top right, bottom right*); David Walker, Lamont-Doherty Geological Observatory (*middle right*)
102	Andrew Christie
103	John A. Whitehead, Jr., Woods Hole Oceanographic Institution
105	Enid Kotschnig
106	Bernd Heinrich (*top*), Bernd Heinrich and George R. Silver (*bottom*)
107	Patricia J. Wynne
108	Bernd Heinrich
109–110	Patricia J. Wynne
112	Lennart Larsen
114	Tom Prentiss
115	T. Douglas Price
116	T. Douglas Price (*top*), Tom Prentiss (*bottom*)
118–119	Tom Prentiss
121	T. Douglas Price
123–126	Michael Goodman

LETTERS

To the Editors:

Although the news article on AIDS ["Science and the Citizen," SCIENTIFIC AMERICAN, January] was a fair description of AIDS in the developed world, it fell into the same trap as most commentaries do by tossing a sidelong glance at the disastrous situation in Africa. It is imperative that we realize what the disease means to that continent. In countries where even measles cannot be controlled—because of endless problems related to infrastructural shortfalls—the presence among people of childbearing age of a disease with a long latency period, transmitted by blood and sexual contact, has dreadful long-term consequences.

The World Health Organization estimates that a million people are already dead or dying of AIDS in Africa. Many of the people on whom the affected countries depend for their labor force are relatively mobile, live in cities and have largely put aside tribal customs; hence they are now very likely to be exposed to the virus. As a consequence banks, hospitals, schools and middle-management cadres are being depleted of their valuable professionals.

The unscreened blood supplies and unsanitary medical practices—in particular the use of unsterile needles—common in these nations are transmitting AIDS at a rapid rate. The cost of preliminary blood screening with the ELISA test (which may not be a good indicator of infectivity because of new strains of the AIDS virus in Africa and the possibility of false positive results due to cross-reactivity with malaria) is about equal to what most African countries spend per year for each citizen for *all* health services. In other words, these countries simply lack the money to prevent the transmission of AIDS through the blood supply.

Present WHO efforts toward universal routine vaccination programs in developing countries may have to be put on hold until needle sterility can be ensured. Moreover, many thousands of the infants targeted for vaccination will doubtless be infected with the AIDS virus. What effect will live and attenuated pathogens have on these immune-deficient babies?

As AIDS spreads through the countryside millions of men and women of childbearing age are becoming infected. One can clearly foresee the consequence if extensive health-education measures are not instituted soon. Within 10 years the central part of Africa, from the Atlantic to the Indian Ocean, may be peopled by the elderly and the very young. Access to the populations and the availability of funds will force a choice about the focus of AIDS-control efforts. As a former health educator in Africa, I suggest that priority be given to school-age children who have not yet become sexually active.

Even if a vaccine and/or a curative treatment were found tomorrow, another hurdle would remain: success in the delivery of health care in Africa is marginal at best. Ordinary childhood diseases kill hundreds of thousands of Third World children annually because it is impossible to get around woefully inadequate infrastructures. The insidious nature of AIDS and its long latency period, coupled with the fact that rural populations are largely illiterate and have widely varying cultures and hundreds of languages, make the prevention of AIDS a problem of dumbfounding complexity. We simply cannot afford to view the situation in Africa as analogous to that in developed countries. The disease presents a special problem on that continent, one that sooner or later will heavily affect the entire world economy. The human misery created by this disease is dreadful to contemplate. It is as if the Sahara were rolling down across central Africa, sucking the life from these nations.

ANN GIUDICI FETTNER

Washington, D.C.

To the Editors:

Scientific American ["Science and the Citizen," December, 1986] errs in implying that Nobelist Dudley R. Herschbach was the first to use crossed molecular beams successfully to study the details of a chemical reaction. That honor belongs to Ellison H. Taylor and Sheldon Datz of the Oak Ridge National Laboratory, who published the results of the first successful experiment in the field in 1955. Herschbach, in fact, acknowledged Taylor and Datz's priority in his 1961 article in the *Transactions of the Faraday Society* with these words: "The feasibility of such experiments was established in 1955 by the work of Taylor and Datz on the reaction $K + HBr = H + KBr$."

The 1955 paper is widely regarded as a classic and as the beginning of the modern study of the dynamics of chemical reactions. It showed that one could study chemical reactions under ideal conditions: single collisions, separate control of reactant energies and angular analysis of the products. It also indicated a class of reactions that were practicable to study in this way, and it described the new detector the authors had invented, which made the work possible. In subsequent papers Datz, Taylor and various colleagues studied other reactions, investigated in detail the properties of their two-filament detector, laid out the calculations necessary for interpretation, reviewed the entire field of chemical applications of molecular beams and pointed out directions for further elaboration of the method.

The possibilities suggested by this work generated intense interest and brought many workers into the field, but for more than a decade almost all the work was based on the original class of reactions and on the two-filament detector of Taylor and Datz.

The fundamental difficulty in doing chemical-beam experiments had always been the minute numbers of product molecules they produce and the lack of a detector combining high sensitivity with the ability to distinguish products from elastically scattered reactants. Taylor and Datz originally planned to solve the difficulty by combining neutron activation of bromine with surface-ionization detection of both K and KBr on the usual tungsten filament. That intention explained their presence at Oak Ridge, a nuclear laboratory, but they hoped to avoid the drudgery inherent in activation measurements by finding some modification of the Langmuir gauge that would provide differential capability. They discovered that a platinum filament was insensitive to combined potassium (KBr in this case) and then constructed the successful two-filament gauge.

This letter is not intended to detract in any way from the recognition of the brilliant work of Herschbach and his fellow Nobelists, Yuan T. Lee and John C. Polanyi, but simply to ensure proper credit for earlier accomplishments. In a broader context, I think it is important that this contribution from a national laboratory be widely understood. The funding at national laboratories of basic research not directly related to practical missions is sometimes criticized as supporting rather pedestrian efforts. As this example shows, breakthroughs can appear wherever imaginative scientists have a stimulating environment.

ALVIN M. WEINBERG

Institute for Energy Analysis
Oak Ridge Associated Universities
Oak Ridge, Tenn.

IT WAS A MOMENT YOU STRUGGLED FOR. ARGUED FOR. WORKED FOR.
A MOMENT OF COMMON GOALS AND AGREEMENT. OMEGA. FOR ALL YOUR SIGNIFICANT MOMENTS.

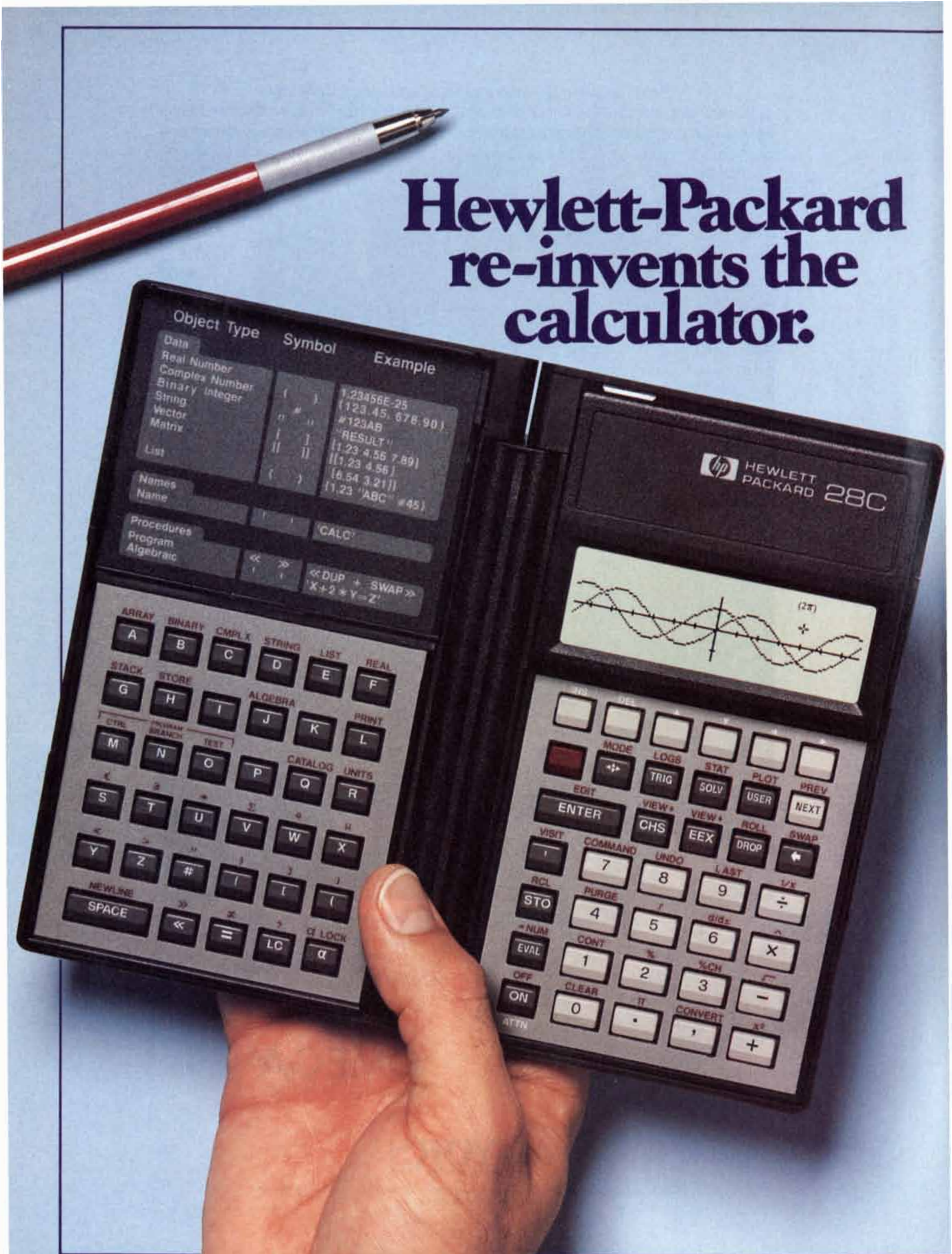
Registered Model

OMEGA ALWAYS MARKS SIGNIFICANT MOMENTS.
IN THE OLYMPICS. IN THE SPACE PROGRAM.
IN SIGNIFICANT LIVES LIKE YOURS.
THE OMEGA CONSTELLATION. ADVANCED SWISS TECHNOLOGY. WATER RESISTANT. IN STAINLESS STEEL AND 18K GOLD.

Ω
OMEGA

MAYOR'S
Fine Jewelers since 1910.

Miami · N. Miami · Coral Gables · Hialeah · Hollywood · Plantation · Ft. Lauderdale · Coral Springs · Pompano
Boca Raton Boynton Beach · Orlando


Hewlett-Packard
re-invents the
calculator.
Object Type
Symbol
Example
HEWLETT
PACKARD 28C

50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

MARCH, 1937: "The American Society of Mechanical Engineers, the Institute of Aeronautical Sciences and the Society of Automobile Engineers recently combined in organizing a symposium titled 'The Next Five Years in Aviation.' Among the opinions advanced were that air-cooled engines will develop well over 1,000 horsepower in the near future and that the fuel consumption of the gasoline aircraft engine will be reduced to .35 pound per horsepower-hour as compared with .5 pound today. Igor Sikorsky spoke with his customary vigor and authority about aircraft weighing 100,000 to 200,000 pounds. Cruising speeds will be increased to 200 miles an hour for flying boats and 250 miles an hour for land planes."

"The peacetime applications of flying never cease to increase in scope and number, as a good offset to the dangerous functions of the airplane in war. The New Jersey Fish and Game Commission has saved the lives of thousands of ducks in the lower bays of both New Jersey and New York by spreading grain from the sky at a time when ice prevented approach to the starving birds by any other means."

"The main highways of the future are pictured by Thomas H. MacDonald, chief of the United States Bureau of Public Roads, as broad, unobstructed surfaces over which traffic can flow smoothly with safety and comfort. 'At the speeds we now foresee,' he said, 'we will want a 22-foot road for two-lane traffic. Its surface will be consistently smooth and non-skid, and it will be so designed that at no place will the traveler suddenly encounter the unexpected. Where the traffic is so heavy that two lanes will not carry it, there will be four-lane roads consisting of two lanes on each side of a center parkway.'"

"Development at the Massachusetts Institute of Technology of a compact magnet capable of producing the highest permanent magnetic field ever attained has opened the door to a new and significant field of exploration in the world of matter. The new magnet was designed by Dr. Francis Bitter of the department of mining and metallurgy. The magnetic field produced in its first test was 150,000 times more intense than the earth's field."

"It may sometimes be unscientific to be too 'scientific' about the food given to infants. An article in the *Journal of the American Medical Association* described experiments in which babies were daily surrounded with a wide variety of foods, some of them not usually regarded as suitable for infants, and allowed to grab at will. In many cases the infants chose and ate immense amounts of some one food, and exactly *nothing* happened. The experiments tended to bolster the theory that infants, like other animals, instinctively know what is good for them and eat it if they can get it."

"The addition of small amounts of salt, sodium sulfate or sodium phosphate to soap, the alkalinity of which has already been adjusted to give the best results, improves its washing effectiveness."

SCIENTIFIC AMERICAN

MARCH, 1887: "Inventors often complain of the difficulty experienced in inducing capitalists to join them in their enterprises. Not infrequently the blame rests as much with the inventor as with the man of money. The capitalist is often blamed for not seeing into the advantages of an enterprise, when the fact is it has never been presented to him in the right light. Every man, therefore, who would seek the aid of capital in furthering his plans for introducing an invention should first be prepared to show the whole state of the art covered by such an invention, and wherein the improvement lies. Second, he should, if possible, show what particular market needs to be supplied with such improvement, and something approximating to the returns which reasonably may be expected. Third, he should have some well settled plan of introducing the new product or furthering the new scheme. If his invention is worth pushing, in nine cases out of ten there will be little trouble in procuring financial help if the proper methods be employed."

"When the last census was taken, to wit in 1880, the census man did not consider the electric lighting investment of sufficient importance to warrant him in collecting the data. Today the investment is $107,000,000. And how long has it taken for this vast sum to be attracted to the electric lighting field? Only six years!"

"In the house of Thomas A. Edison is a remarkable memento of Henry Ward Beecher. The inventor's phonograph for impressing on a soft metal sheet the utterances of the human voice, and then emitting it again by the turning of a crank, has never been put to any very valuable use. But he utilized it to make a collection of famous voices. Edison is probably the only man who can revive the silenced voice of the great preacher."

"At a meeting of the Caucasian Medical Society, Dr. A. P. Astvatzaturoff of Tiflis drew attention to the danger of infection arising from the promiscuous use of the mouthpieces of public telephones. To prevent any accident of the kind, he recommends that the mouthpiece should be disinfected every time after or, still better, before it is used. In other words, some disinfectant fluid should be kept at every telephone station and the speaker should dip the mouthpiece into it."

"In the SCIENTIFIC AMERICAN of November 14, 1885, we gave some interesting illustrations of sail skating as practiced in the vicinity of Copenhagen, and the enlivening sport which is to be had out of the ice boat, as the swift craft are managed by Hudson River ice yachtmen, has often been described; but the illustration herewith represents a decidedly new phase of exhilarating enjoyment in diversions of this kind. It is a practical snow yacht, constructed and used during the past winter by Dr. H. M. Wheeler of Grand Forks, Dakota, our illustration being made from photographs he has furnished us."

A snow yacht on the Dakota prairie


LF
180

THE AUTHORS

FRANK R. VELLUTINO ("Dyslexia") is professor and research psychologist at the State University of New York at Albany, where he is also director of the Child Research and Study Center. He has been on the Albany faculty since 1966; before that he worked for two years as a clinical psychologist at the Veterans Administration Hospital in Brockton, Mass. He earned both his master's and his Ph.D. in psychology, the latter in 1964, from the Catholic University of America. Most of the research Vellutino cites in his article was funded by the National Institute of Child Health and Human Development.

JAMES M. HOGLE, MARIE CHOW and DAVID J. FILMAN ("The Structure of Poliovirus") were awarded the 1985 Newcomb Cleveland Prize by the American Association for the Advancement of Science for the paper in which they originally reported their findings. Hogle, a member of the department of molecular biology at the Research Institute of Scripps Clinic, developed an enthusiasm for viruses in the course of his doctoral research at the University of Wisconsin at Madison. He got his Ph.D. in 1978 and did postdoctoral work at Harvard University. Hogle has been at Scripps since 1982. Chow met Hogle six years ago while doing postdoctoral research at the Massachusetts Institute of Technology. Currently a member of the applied biological sciences department at M.I.T., she received her Ph.D. from Yale University in 1981. Before joining the university's faculty Chow worked at the Whitehead Institute for Biomedical Research and at the Center for Cancer Research. Filman, who says his "native language is Fortran," joined the research effort when Hogle arrived at Scripps. He has been programming computers since his high school years and studying crystallography since 1975. Filman earned his Ph.D. from the University of California at San Diego in 1981 and did research there until 1983, when he moved to Scripps.

WILLIAM D. PHILLIPS and HAROLD J. METCALF ("Cooling and Trapping Atoms") began their collaboration in 1981 when Metcalf first served as a consultant to the National Bureau of Standards. Phillips had been at the bureau since 1978; before that he got his Ph.D. from the Massachusetts Institute of Technology in 1976 and stayed on for two years of postdoctoral research as a Chaim Weizmann Fellow. Phillips is a fellow of the American Physical Society and an active participant in his church's gospel choir. Metcalf went to Brown University for his Ph.D. in physics, awarded in 1967, and spent a year at Brown as a research associate. In 1968 he moved to the State University of New York at Stony Brook, where he is professor of physics. Metcalf has also taught outside the U.S.; he is currently a visiting professor at the École Normale Supérieure in Paris.

J. F. POWER and R. F. FOLLETT ("Monoculture") are both veteran soil scientists with the Agricultural Research Service of the U.S. Department of Agriculture. Power has worked for the service since receiving his Ph.D. from Michigan State University in 1954. After six years at the service's extension in Sidney, Mont., he moved to its branch in Mandan, N.D., and then to the branch in Lincoln, Neb., in 1979. He is still at Lincoln, as research leader and is also adjunct professor at the University of Nebraska. Power has been a national technical adviser for both soil management and land reclamation. Follett leads research on soil-plant nutrients for the Agricultural Research Service extension at Fort Collins, Colo., where he has worked since 1981. In 1976 he joined the service's headquarters in Beltsville, Md., and became a national program leader, a position he retained until 1985. He has also had research assignments in Mandan, N.D., and Ithaca, N.Y. Follett, whose Ph.D. from Purdue University was granted in 1966, has edited two books. He received the Agriculture Department's distinguished service award in 1984.

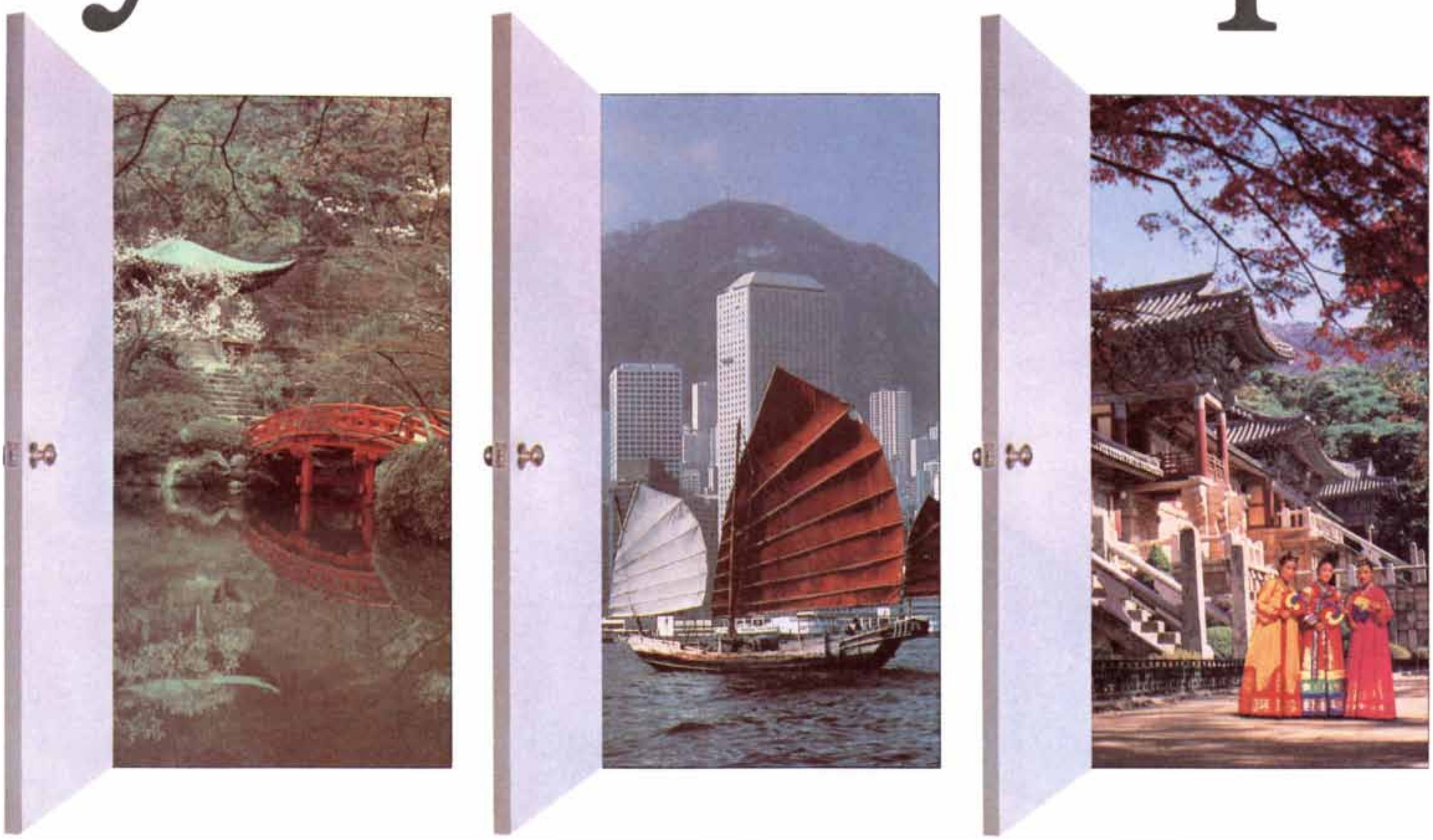
YASER S. ABU-MOSTAFA and DEMETRI PSALTIS ("Optical Neural Computers") are both members of the faculty at the California Institute of Technology. Abu-Mostafa, assistant professor of electrical engineering and computer science, got a B.Sc. in 1979 at Cairo University and an M.S.E.E. in 1981 from the Georgia Institute of Technology. He has been on Caltech's faculty since completing his Ph.D. there in 1983. Abu-Mostafa's doctoral research was honored as the most original Caltech thesis for that academic year. Psaltis earned his B.Sc., M.Sc. and Ph.D. from Carnegie-Mellon University. After getting his doctoral degree in 1977, he remained at the university as a research associate and then as visiting assistant professor. He has been in Caltech's electrical engineering department since 1980. Psaltis has published more than 100 articles on topics such as pattern recognition and optical devices.

ENRICO BONATTI ("The Rifting of Continents") says his years of fieldwork have presented "one delightful problem": how to combine passive contemplation of the earth's beauty with analytical observation of its processes. Bonatti holds a *laurea,* awarded in 1960, from the Scuola Normale Superiore in Pisa. In 1959 he attended Yale University on a Fulbright scholarship and in 1960 he went to the Scripps Institution of Oceanography. Four years later he moved to the University of Miami. Bonatti has held his present position as senior research scientist at Columbia University's Lamont-Doherty Geological Observatory since 1976.

BERND HEINRICH ("Thermoregulation in Winter Moths"), professor of zoology at the University of Vermont, has been studying thermoregulation in insects for almost 20 years. After receiving bachelor's and master's degrees from the University of Maine, he went to the University of California at Los Angeles, which granted his Ph.D. in 1970. He spent the next 10 years in the entomology department of the University of California at Berkeley before moving back east to Vermont. Heinrich is now examining the foraging habits of ravens but maintains that he has not changed fields: "I simply use different organisms to study the same questions."

T. DOUGLAS PRICE and ERIK BRINCH PETERSEN ("A Mesolithic Camp in Denmark") have been able to explore Scandinavia's Mesolithic profile together in spite of the distance separating their home stations. Price is chairman of the anthropology department at the University of Wisconsin at Madison, where he has been teaching since 1974. He got his Ph.D. in 1975 from the University of Michigan, and he has directed archaeological projects in the Netherlands and Denmark. Price is now studying bone samples for evidence of the prehistoric human diet. Petersen's investigations have taken him to Iran and Greenland as well as southern Scandinavia. He is a lecturer at the Institute of Prehistoric Archaeology at the University of Copenhagen, where he earned his *magister* in 1970. An affiliate of the National Museum of Denmark, Petersen has also been involved in designing a new archaeology museum in Søllerød.

Look who puts Asia at your doorstep.



No matter where you are in the U.S., we make it easy to get to Hong Kong, Tokyo, Seoul, Manila, Taipei, Osaka, Shanghai, Guam or Okinawa.

We offer you hassle-free, one-airline service to Asia from over 200 cities. With nonstops from New York, Chicago, Seattle, San Francisco, Los Angeles and Honolulu.

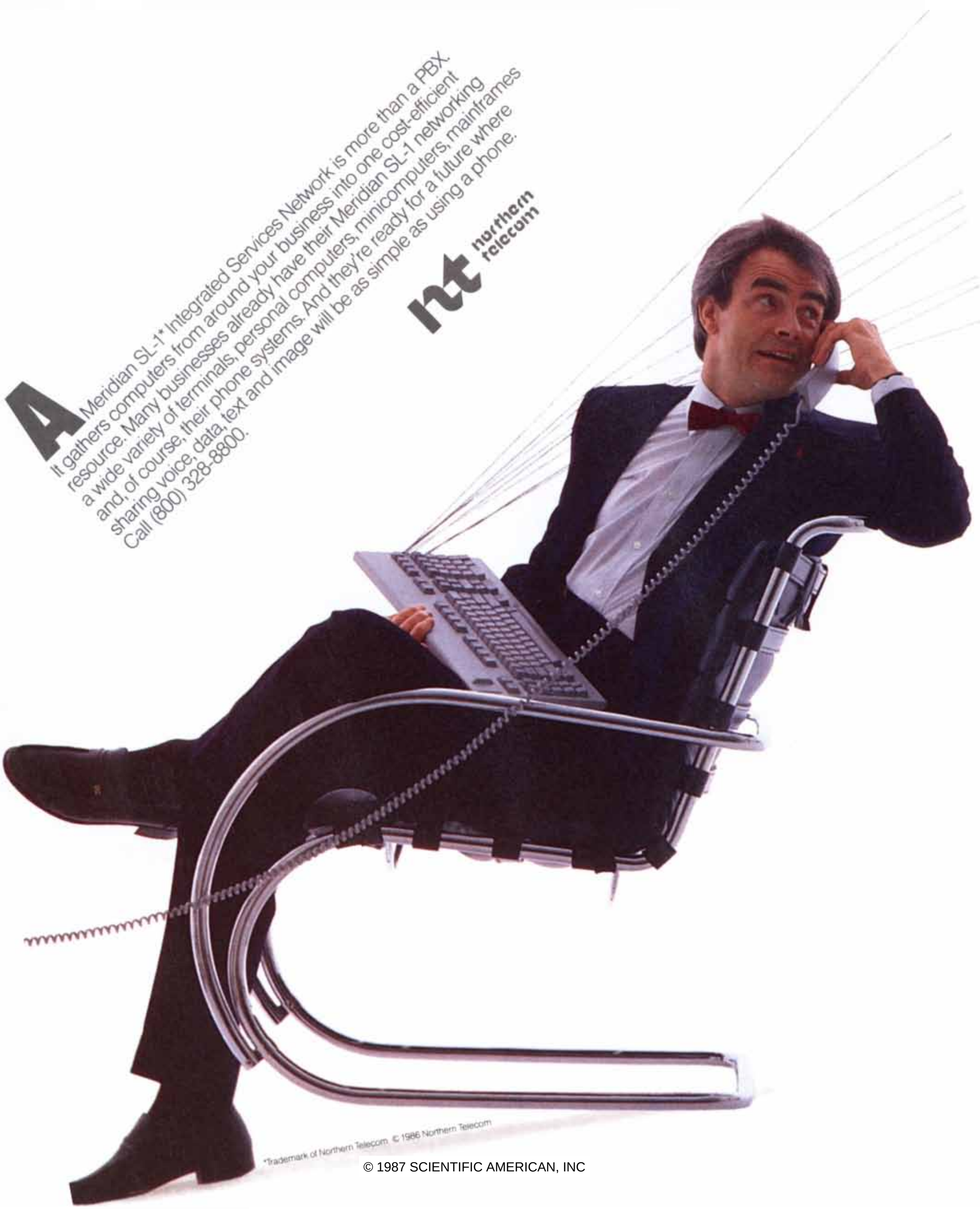
And we'll fly you to Asia in nothing short of wide-cabin 747 comfort.

With 40 years of transpacific experience, we make distant lands seem a lot closer.

Call your travel agent or call Northwest at (800) 447-4747 for international reservations.

Look to us.

NORTHWEST


A Meridian SL-1* Integrated Services Network is more than a PBX.
It gathers computers from around your business into one cost-efficient
resource. Many businesses already have their Meridian SL-1 networking
a wide variety of terminals, personal computers, minicomputers, mainframes
and, of course, their phone systems. And they're ready for a future where
sharing voice, data, text and image will be as simple as using a phone.
Call (800) 328-8800.
nt northern telecom
*Trademark of Northern Telecom. © 1986 Northern Telecom

meridian sl-1

NETWORKING

COMPUTER RECREATIONS

Braitenberg memoirs: vehicles for probing behavior roam a dark plain marked by lights

by A. K. Dewdney

A vast plain hums with the activity of hundreds of toy vehicles. Some cluster worshipfully at the base of giant light bulbs. Others seek darker places or hover uncertainly between light and shade. What land is this? What madness?

It could be called the land of synthetic psychology, a gedanken ground invented by Valentino Braitenberg of the Max Planck Institute for Biological Cybernetics in Tübingen. Braitenberg's thesis that biological behavior is easier to synthesize than to analyze is aptly illustrated by the vehicles that inhabit the great plain. Employing only elementary mechanical and electrical devices, even the simplest control circuits give rise to behavior that Braitenberg calls love, aggression, fear and foresight. A complete description of Braitenberg's thought experiment can be found in his handy little book titled *Vehicles: Experiments in Synthetic Psychology*.

Before exploring the world of Braitenberg's vehicles, there is a story of my own that illustrates his thesis. A certain professor of computer science once brought a curious object into his artificial-intelligence classroom. He displayed a golden sphere that he claimed (tongue in cheek) had fallen from a UFO that very morning as he was on his way to the campus. Turning his back to the class, he made a few adjustments and the sphere began to buzz. When he placed the sphere on the smooth classroom floor, it abruptly began to roll by itself. Whenever it bumped into an obstacle such as a chair leg, it would pause for a moment and then roll around it. At one point the sphere even rolled behind the door that stood open against one wall. The sphere buzzed and bumped mysteriously for a minute and suddenly re-emerged from the cul-de-sac as though satisfied with its explorations. Readers might enjoy pondering what secret force controls the sphere. How does the sphere avoid obstacles in the way I have described? The answer will be revealed next month.

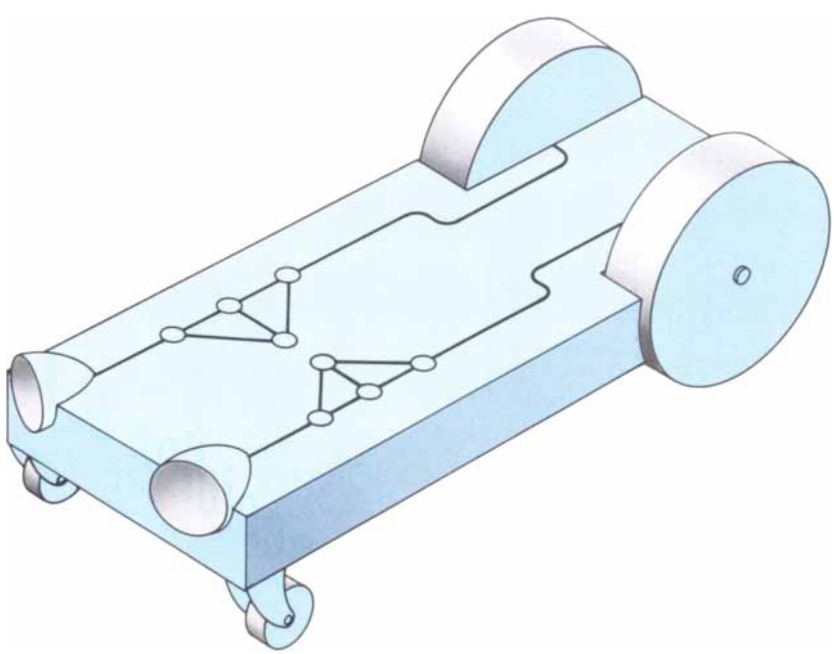
Braitenberg's vehicles share with the golden ball the characteristic that only the initiated know how they work; others can only shake their heads in puzzlement. In what follows I have taken some liberties in simplifying and formalizing Braitenberg's game so that readers can join the initiated and engage in the adventure.

A vehicle can be visualized as a toy-size schematic automobile with two independent driving wheels at the back and two slave wheels (like casters) at the front [*see illustration on this page*]. The difference in turning rates of the back wheels determines where the vehicle will go.

The wheels are attached to the simplest body possible, namely a rectangular slab of material. The vehicle designer can fasten various sensors around the edge of the slab. A circuit that connects the sensors to the rear wheels completes the vehicle. The wiring is mounted on the top of the slab somewhat in the manner of a printed circuit. Only one piece of optional equipment is available for the vehicles, a rectangular cover that can be placed over the circuit. The cover is not meant to protect the vehicular brain from the rain (the land of synthetic psychology has no weather) but to shield it from the prying eyes of outsiders; we have no wish to reveal how simple the little circuits actually are. Outsiders are therefore free to draw the most fantastic conclusions about what the vehicles are up to and just how they work. We do not want to spoil their fun.

The standard vehicle has two wide-angle "eyes," or photoreceptors, attached to the front of the slab. Each receptor is aimed slightly to the outside of the medial line. A light source ahead and to the left of the vehicle will therefore register more strongly in the left receptor than in the right receptor.

The two receptors are connected by wires to motors that operate the rear wheels. Signals in the wires consist of trains of discrete electric pulses: the more light a given receptor receives, the more pulses it sends along its output wire. At low levels of illumination only a few pulses are sent per second. At high levels many are sent. In both cases we imagine the pulses to be evenly spaced in time. At the other end of the wires the pulse trains are translated into motor commands. The more



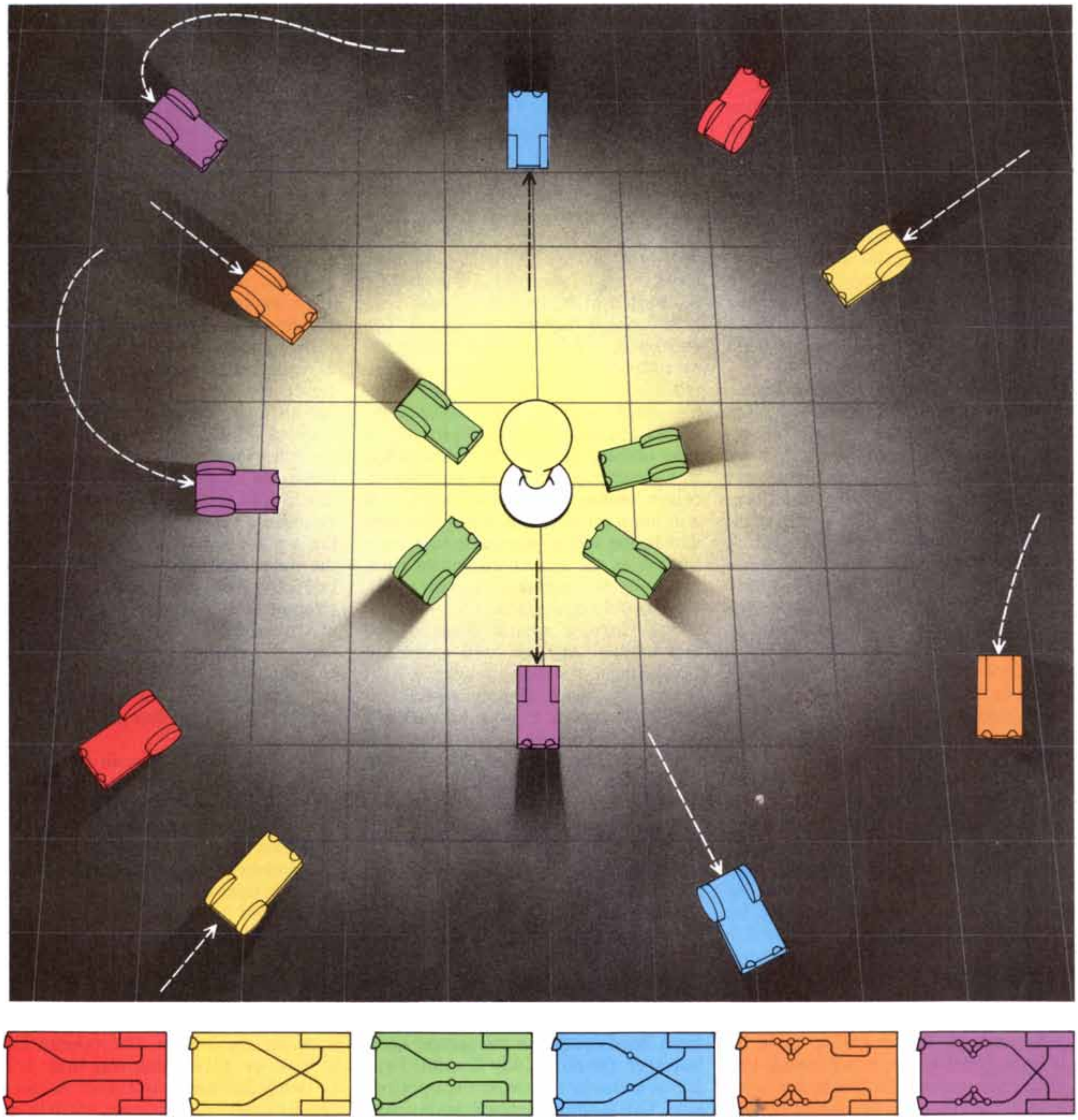
The standard vehicle

pulses that arrive per second at a given wheel, the faster its motor will drive it.

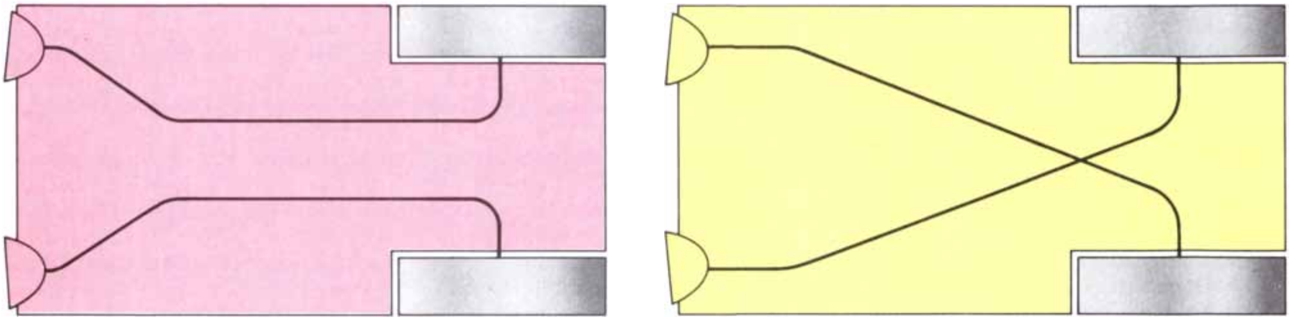
In the simplest vehicles two wires connect the eyes directly to the wheels. Two possibilities arise: the wires may be crossed or uncrossed. In the latter case each wire connects directly to the motor behind it. What happens when such a vehicle is placed in the middle of an infinite, darkling plain punctuated here and there by brightly shining light bulbs? If the vehicle is near one of the bulbs, more or less facing it, the vehicle will rush forward but immediately begin to turn away from the light as though frightened by its brightness. The fearful vehicle will run out into the night, slowing as it goes. In general it will creep slowly about the plain, avoiding bright places. Occasionally it may be forced to run through a gap between two relatively close bulbs. It will speed up like a frightened rabbit, hoping to get over the uncomfortable experience as quickly as possible.

An outside observer might be baffled by the vehicle's behavior, but the explanation is quite simple. If there is a light source ahead of the vehicle and to its left, for example, the left eye will receive more illumination than the right one. Consequently the left motor will run at a higher speed than the one on the right; the vehicle will therefore turn to the right, away from the source of light. The mirror image of this behavior is exhibited when the light source is to the right. The vehicle turns to the left, away from the source. Once the light is out of sight the vehicle immediately slows down, its speed governed only by the average illumination of distant sources in its visual field.

What about the vehicle with crossed connections? Its behavior could hard-



Vehicles roam the plain in the land of synthetic psychology



Direct **(left)** *or crossed* **(right)** *connections make for different behaviors*

ly be more different. Placed on the plain at a great distance from any light bulb, it swings slowly in the direction of the one dominating its visual field. As it draws near, it goes faster and faster. Finally, at top speed, it runs straight into the bulb, smashing it. Would an outside observer not conclude that the vehicle with crossed connections behaves aggressively?

In the absence of any electronic intermediaries between receptors and motors, the two vehicles I have described virtually exhaust the available behavioral repertory. To produce more complicated behavior one can introduce a kind of abstract neuron. I call it a neurode to distinguish it from the real thing.

A neurode is actually a formal computing element that receives pulses from other neurodes or from receptors. The pulses are conducted by wires. A neurode will generate pulses of its own under certain conditions. Somewhere a clock ticks at the arbitrary rate of 100 cycles per second. A neurode will fire at the end of a cycle—and only at the end of the cycle—if the number of pulses it receives from other neurodes during the cycle equals or exceeds a preset number. (A pulse takes somewhat less than a complete cycle to travel between any two neurodes that are directly connected.) The preset number is called the neurode's threshold.

In the most general form of vehicle, pulses are delivered to neurodes by means of two kinds of wire, excitatory wire and inhibitory wire. A neurode will fire if it receives enough pulses along its excitatory wires so that its threshold is met. The neurode will be stopped from firing during a cycle if it receives a pulse along an inhibitory wire during the cycle.

The clock also governs receptors and motors. A receptor may send a pulse only at the end of a cycle. If it fires every cycle, it will send 100 pulses per second, the maximum level of activity a receptor is capable of sustaining. The motors may also receive a maximum of 100 pulses per second from the neurodes that control them. We assume, moreover, that each pulse delivered to a motor turns its wheel by a small angle. Such devices actually exist. They are called, appropriately enough, stepping motors.

By incorporating two neurodes into the hitherto ultrasimple circuits, the behavior of the two vehicles is radically altered: the fearful vehicle becomes a light-bulb lover and the formerly aggressive vehicle becomes shy. Their behavior is not simply reversed, however; the change is subtler. In both cases a single neurode of threshold 0 is inserted in the middle of each wire connecting a receptor to a motor. The section of ordinary wire between the receptors and the neurodes is then replaced with inhibitory wire. As a consequence, when the neurode receives a pulse from the receptor, it will send nothing to the motor attached to it. Conversely, when the neurode receives nothing from the receptor, it will fire a pulse to the motor. In other words, the higher the pulse rate from the eye is, the slower the motor runs; the lower the pulse rate from the eye is, the faster the motor runs.

Now form a mental image of the first vehicle, the one with uncrossed connections. As it approaches a light source, it steers directly toward it, slowing as it goes. Finally it comes to a stop, facing the light bulb in quiet adoration. The second vehicle's crossed connections, on the other hand, now cause it to turn slowly away from a nearby source, then to speed off into the darkest area it can see.

The reasons for the two behaviors should be clear to anyone who knows the circuits. In both cases the greater the illumination of the eye receptors is, the slower the motors run. But in the first case the more fully illumined eye drives a slower motor on the same side. This tends to correct deviations from dead center. In the second case, however, the more fully illumined eye drives a slower motor on the opposite side. This tends to increase deviations.

So far the response of the motors to receptor signals has been either directly proportional or inversely proportional to the pulse rate. By means of a simple little circuit of four neurodes, a nonlinear response can be arranged. The behavior of the four-neurode circuit is quite ordinary as long as the receptor sends fewer than 50 pulses every second to its motor: the more pulses per second from the receptor there are, the faster the motor goes. Beyond 50 pulses per second, however, matters become interesting. The motor begins to slow down: the more pulses per second from the receptor there are, the slower the motor goes.

To explain how the four-neurode circuit works, I must introduce a term called a pulse rate, which is the average number of pulses per cycle. A receptor that fires 50 pulses per second, for instance, has a pulse rate of 1/2, because 50 pulses per second divided by 100 cycles per second equals 1/2 pulse per cycle. (Remember that the clock ticks at 100 cycles per second.) Because a receptor or a neurode can fire a pulse only at the end of a cycle, the pulse rate can never exceed 1. In brief, the four-neurode circuit is constructed so that if an eye sends pulses at a rate less than 1/2, all the pulses get through to the motor. If the pulse rate is greater than 1/2, however, the circuit acts as a net: the more pulses that are sent, the fewer that get through to the motor.

The four-neurode circuit consists of a chain of three neurodes of threshold 1 and an additional neurode of threshold 2 [*see top illustration on page 22*]. The first two threshold-1 neurodes send pulses to the threshold-2 neurode, which in turn sends an inhibitory pulse to the third threshold-1 neurode in the chain. The net effect, so to speak, is to kill off all but the first pulse in any train of consecutive pulses that travel through the miniature plexus. Readers are encouraged to confirm this operation for themselves. The assumption is that for pulse rates less than 1/2 no two pulses are consecutive. At rates greater than 1/2, however, clusters of pulses appear.

Outside observers are invited to

DE V

The Contemporar

At Cadillac, we believe today's luxury cars should reflect today's values. The newly styled, longer De Ville is as contemporary as it is Cadillac. This six-passenger luxury car remembers why you chose Cadillac in the first place with renowned comfort and convenience enhanced by leading-edge technology. De Ville also delivers a smooth, comfortable ride while answering your need for precise driver control. There's even an Anti-Lock Braking System available.

The new look of luxury.

This year, America's most successful luxury car adds the further distinction of new Cadillac styling refinements. The restyled front end with a new, bolder grille and composite headlamps takes on a clean, aerodynamic design. The profile assumes a longer silhouette. And a new taillamp treatment completes the elegant update.

The increasingly rare comfort of full-size room for six.

How can a luxury car this effortless to maneuver in traffic still accommodate six adults in Cadillac comfort? The answer is a transverse-mounted V8 engine with front-wheel drive. By putting the engine sideways, valuable room is passed along to the spacious Cadillac interior.

*See your dealer for the terms and conditions of this new powertrain limited warranty.

LLE

Spirit of Cadillac.

Cadillac touches that anticipate your every want.

The Cadillac name is synonymous with comfort and convenience. In addition to its six-way power driver seat and Electronic Climate Control system with selected temperature display, De Ville also incorporates innovations such as the ability to operate the stereo radio, the available Astroroof and other conveniences up to 10 minutes after you turn off the ignition or until you open a door.

A unique blend of smooth ride and responsive handling.

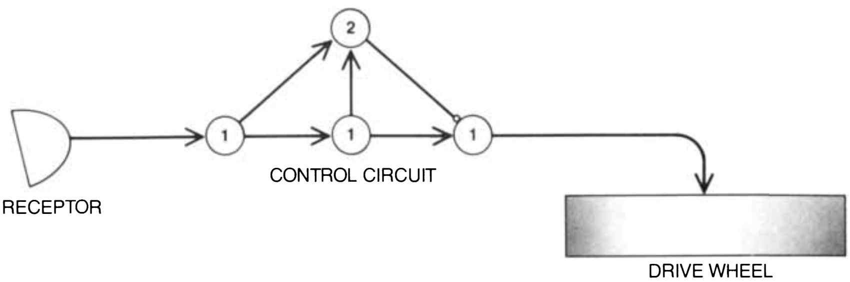
With specifically tuned four-wheel independent suspension and Electronic Level Control, De Ville proves a silky ride and a firm feel-of-the-road are not mutually exclusive. This Cadillac is equally rewarding to driver and passenger. With its available Anti-Lock Braking System, De Ville can help take the panic out of panic stops.

De Ville's new, longer warranty.

Another confidence-inspiring facet of De Ville is its new 6-year/60,000-mile warranty* as well as a 6-year/100,000-mile rust-through perforation warranty.

Your Cadillac dealer invites you to drive the car that defines luxury for today. De Ville, the contemporary spirit of Cadillac.

LET'S GET IT TOGETHER...BUCKLE UP. GM



A nonlinear motor control circuit

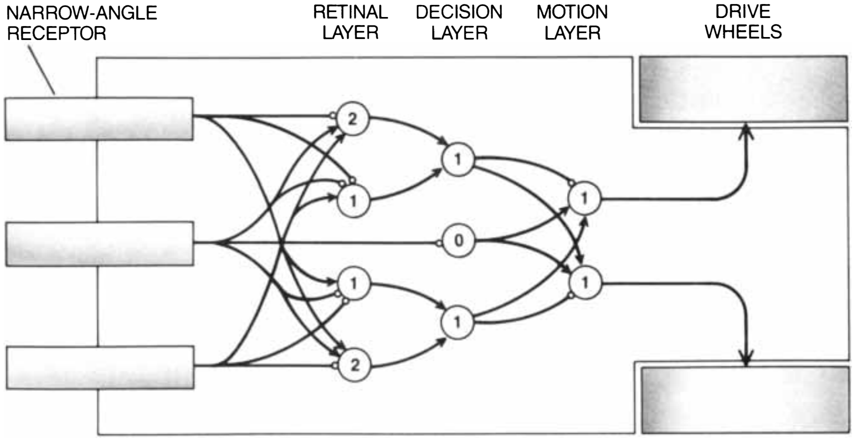
ponder the behavior of the vehicles when each inhibitory receptor wire is replaced by an excitatory wire and each threshold-0 neurode is replaced by a four-neurode nonlinear network. Near the light source, where receptor pulse rates are greater than 1/2, the vehicle with uncrossed eye-to-motor connections behaves like the light-bulb lover. Away from the light source it becomes fearful, shrinking from light of any kind. The vehicle with crossed connections, on the other hand, vacillates. Outside the pulse-rate boundary of 1/2 it is aggressive, heading toward the nearest visible bulb with increased velocity. Within the boundary it becomes hesitant and turns shyly away.

One can readily imagine the great plain populated by all six types of vehicles [*see illustration on page 17*]. Each light bulb would have a handful of quiet worshipers. Farther away vacillating vehicles would follow complicated orbits of indecision. In the outer darkness, meanwhile, the shy and fearful vehicles would creep quietly or whiz anxiously in complicated patterns. Occasionally the devotions of the quiet vehicles would be shattered by the arrival of a violent vehicle. If the process destroyed a bulb, all the vehicles would race off in search of another light source. The instigator of the destruction (assuming it survived the crash in perfect shape) would move off more slowly than the worshipful cars.

For vehicles equipped with neurodes the sky is the limit, behaviorally speaking. Indeed, one can build a perfectly good computer from interconnecting neurodes. But what good is a powerful brain with such limited sensory inputs? To add spice, vehicles could be equipped with enhanced vision, hearing and touch. The actual senses added are largely irrelevant; one would only like to construct a reasonably interesting and varied environment. Enhanced vision involves not only the wide-angle receptors already used but also narrow ones that can be assembled into compound arrays. Visual receptors could also have filters placed over them so that the vehicles could distinguish colors and detect heat (which is infrared radiation). Sound receptors might include primitive ears tuned to various frequency ranges. For touch I visualize long, whiplash antennas extending well out in front of the vehicles.

One could then add buzzers and heat sources to the plain. Vehicles themselves might give off heat. Some light bulbs might buzz.

While I was reading Braitenberg's book an arresting passage caught my eye. In reference to the superior abilities of neurode-equipped vehicles he writes: "You can already guess some of the things that a vehicle fitted with this sort of brain can do, but you will still be surprised when you see it in action. The vehicle may sit there for hours and then suddenly stir when it sights an olive green vehicle that buzzes at a certain frequency and never moves faster than 5 cm/sec."



A predatory vehicle

The description set me to imagining a species of predatory vehicle that would attack any and all vehicles that happened to cross its visual field [*see bottom illustration on this page*]. I have designed only the portion of the predator's brain responsible for pursuit; I offer my design as a sample exercise. Readers might enjoy designing a brain for a behavior that intrigues them.

The predatory brain is made up of three levels of interconnected neurodes. The first level amounts to a retina that sifts meaningful data from the possibilities presented by three narrow-angle visual receptors. The second level decides whether prey is present ahead. The third level consists of two motor neurodes that command the drive wheels.

In detail, there are four retinal neurodes, each of which watches for a specific pattern of light and dark falling on the three photoreceptors. The two neurodes on the left side of the circuit will fire if a dark object (a vehicle by definition) falls within the range of the leftmost one or two receptors. Those two neurodes are connected by excitatory wires to a single threshold-1 neurode positioned at the second level of the vehicle's brain.

The second-level neurode fires at the end of a given cycle if either of the two retinal neurodes fired in the preceding cycle. It therefore fires if a vehicle was spotted on the left side of the visual field. The second-level neurode simultaneously transmits an excitatory pulse to the motor neurode commanding the right drive wheel and an inhibitory pulse to the motor neurode commanding the left drive wheel. The vehicle begins turning to the left so that the target will tend to be centered on the retina. The neurodes on the right side of the predatory brain function in a mirror-image fashion to those on the left. A central, threshold-0 neurode at the second level is connected by a single inhibitory wire to the central retinal receptor and by an excitatory wire to each of the two motor neurodes. The threshold-0 neurode will drive the vehicle straight ahead provided neither of its companions at the second level is also firing, since each of them will inhibit one of the motor neurodes from responding.

Admittedly it is much easier to synthesize one of Braitenberg's vehicles

than it is to analyze the inner workings of a living nervous system, even the primitive one found in the large marine snail *Aplysia* [see "Small Systems of Neurons," by Eric R. Kandel; SCIENTIFIC AMERICAN, September, 1979]. Undoubtedly observers familiar with the rules of the synthetic-psychology game could understand many of Braitenberg's vehicles simply by synthesizing ones of their own. At the same time Braitenberg's vehicles teach us that even the most primitive nervous systems may be capable of behavior that seems complicated or surprising. In any event, neurophysiologists have their hands full with creatures whose complexity is several orders of magnitude greater than that of the vehicles described here.

Michael A. Arbib is a computer scientist and brain theorist who recently moved to the University of Southern California to help develop a program of studies in neural, informational and behavioral science. Arbib's "vehicle" for brain research is a computer-simulated frog called *Rana computatrix*. The frog brain is at present only partially implemented: there is a crude retina, an advanced tectum and as yet little else. Already, however, the frog hops around barriers and distinguishes prey from inanimate objects. It displays simple learning.

According to Arbib, there is something of a resurgence of interest in neural models both as simulations of living creatures and as potential computers. In the case of the former a number of workers have constructed computer simulations of isolated neural systems in simple animals such as locusts, frogs and sea slugs. The simulations are already leading to testable predictions. In the case of the latter, a small number of engineers are investigating formal neurons as elements in parallel computers.

The gulf between synthesis and analysis does not imply that Braitenberg's vehicles should be viewed merely as frivolous playthings. Perhaps more germane than the current gulf between the two is the tendency of some observers to describe the actions of the vehicles in terms of human behavior, to say the vehicles love or hate based on behavior that appears complicated but really is not to those who understand it. Braitenberg seems to imply that the phenomena of love and hate are manifested by human systems as definite in principle as vehicular systems are; as one ascends a scale of complexity, passing from abstract to real creatures in the process, one might be forced to conclude that love and hate were present almost from the beginning.

Braitenberg illuminates the topic in his *Vehicles* by carefully constructing a sequence of machines ranging from simple to complex. At the end of the book he explains how each assumption, behavioral or mechanical, is grounded in real observations. The ability of an array of narrow-angle visual receptors to distinguish among objects might ultimately account for how a fly can choose one's nose to land on in a crowded room.

I have been charged with incompetence by William J. Slattery of Jamestown, R.I. Last December's column on Pixar, the California-based graphics computer firm, featured a description of programs for drawing mountains and plants. What appears to be a fractal mountain drawn by the method of triangular subdivision is shown on page 16 of that issue. According to Slattery, I did not have the wit to recognize Sherlock Holmes's hat when I saw it. For misleading readers I may even be apprehended by the great detective himself.

John McCarthy, an eminent computer scientist at Stanford University, has a more serious comment on the concept of fractal mountains. He notes that real mountains, which are formed by erosion, have many well-developed river valleys. Fractal mountains, on the other hand, do not share such features and would, in fact, be full of lakes. Is the resemblance of fractal mountains to real mountains merely an optical illusion? McCarthy warns against the acceptance of fractal objects as models for real ones without further investigation.

The cross-sectional object I called Mount Mandelbrot might not be a single mountain but an entire range of mountains. Graphic output sent to me by Ken Wright of Grayling, Mich., bears out this conclusion. One way back to singular orogeny is to take some care with the random deflection in the line segments that make up Mount Mandelbrot's profile; if the deflection is biased in the upward direction, rather than being allowed to range up or down with equal probability, one may see a more impressive peak develop.

The program outlined for Mount Mandelbrot wasted some memory by requiring two arrays to hold alternate generations of points constituting successive subdivisions of the mountain's profile. Richard E. Lang of Lincoln Center, Mass., finds a single array is sufficient. Essentially the same algorithm I suggested can be easily converted to make use of the single-profile array as the basis for its own subdivision, so to speak.

The Tandy® 3000 HL makes 286 technology affordable.

We beat the competition . . .

If you're using MS-DOS® based computers and want to expand with more power for less, take a good look at the Tandy 3000 HL. It beats out the IBM® PC/XT-286 in price, performance and choice of options.

on performance . . .

Operating at 8 MHz (vs. 6 MHz for IBM's 286), the 3000 HL's advanced, 16-bit microprocessor delivers up to seven times the speed of a standard PC's microprocessor. This means you can run the business software you've grown accustomed to faster than ever.

on flexibility . . .

And you can choose your options. Expand with more memory, hard disk drives, communications and more. And you can connect the 3000 HL with other MS-DOS based computers with the ViaNet™ Local Area Network.

and on price!

The floppy-based Tandy 3000 HL starts at only $1699. The IBM PC/XT-286 costs $3995*. *And a comparably equipped Tandy 3000 HL with a 20-megabyte hard disk and other options still costs less than IBM's 286.*

Compare the Tandy 3000 HL. It delivers today's advanced computer technology . . . for less.

Send me a Tandy 3000 HL brochure.

Mail To: Radio Shack, Dept. 87-A-502
300 One Tandy Center, Fort Worth, TX 76102

Name ____________________
Company ____________________
Address ____________________
City ____________ State ______
ZIP ________ Phone ____________

Available at over 1200 Radio Shack Computer Centers and at participating Radio Shack stores and dealers.

TANDY COMPUTERS: In Business . . . for Business™

Radio Shack® COMPUTER CENTERS
A DIVISION OF TANDY CORPORATION

*Based on IBM price list as of Sept. 2, 1986. Tandy 3000 HL price applies at Radio Shack Computer Centers and participating stores and dealers. Monitor sold separately. MS-DOS/Reg. TM Microsoft Corp. IBM/Reg. TM International Business Machines Corp. ViaNet/TM ViaNetix, Inc.

BOOKS

High culture in the caves, talking hands, was Einstein really right, life at the top

by Philip Morrison

DARK CAVES, BRIGHT VISIONS: LIFE IN ICE AGE EUROPE, by Randall White. The American Museum of Natural History and W. W. Norton & Company, Inc. ($35).

Dark, sealed caves preserve what they hold, and certainly they draw the archaeologists. These days it becomes increasingly clear that the people of Ice Age Europe did not occupy confining natural shelters in the limestone for want of inventiveness. They chose how to live: three-fourths of all their sites are south-facing, 90 percent are near water. Open-air sites, often commanding a vista over the sweet green meadows of the Périgord, were favored by those predecessors of ours. The archaeologists are beginning to search more widely; the author's own photograph persuades as it invites us to a living floor he found as a test of his topographic analysis on the rolling uplands of the Vézère River. Far from sheltering rocks, the cunningly made tools are in place after 20,000 years, buried a few feet in the mute soil.

Another photograph shows 100 river cobblestones brought to form a foundation floor for a skin or wood shelter. Those stones were fire-heated before they were assembled, perhaps to be set hot into the frozen ground. Such an inlaid floor would not leave its builders in a messy and uncomfortable pool of mud once the frozen ground had thawed within the hearth-warmed shelter. Flint for tools also was heated to a "porcelainlike sheen" before pressure flaking. Cobbled fireplaces and stone-filled pits for boiling food are shown as well.

An even more remarkable shelter is shown in a color photograph made at Mezirich beside the Dnieper River in the Ukraine. The double spread displays the excavation of an impressive house wall formed by careful herringbone stacking of scores of the big lower jawbones of woolly mammoths. The builders of these grotesque dwellings 15,000 years ago were a band of some 25 people; this shelter was assembled not from their own kill of game but from the skeletons of mammoths long dead, found strewn in the neighborhood [see "Mammoth-Bone Dwellings on the Russian Plain," by Mikhail I. Gladkih, Ninelj L. Kornietz and Olga Soffer; SCIENTIFIC AMERICAN, November, 1984]. Such a bone hut could be built in a week by 10 people working together. Another Ukrainian site disclosed what seems to be a "stone age orchestra," its pride a sonorous osteophone assembled from mammoth bones.

The fresh and critical essay on the rise of *Homo sapiens* in Europe informs this beautifully illustrated survey. The book includes a checklist of the 200 or so pieces of bone and stone exhibited in New York: the splendid portable art of "the first of the world's great civilizations."

The chapters tell briefly of the coming and survival of the people who endowed southwestern France (the majority of the material here is French, although the view taken is never provincial) with their novel art, architecture and artifacts. They came at the end of the long, conservative epoch of the Neanderthals, those earlier artisans who "made beautiful tools stupidly" and almost without change for more than 60,000 years. Now we try to read social organization—population, social divisions, trade—from the complex burials and relics of the Upper Paleolithic, our conspecific forebears.

There are an overwhelming number of wonderful objects. A limestone bas-relief is typical. It shows a pair of wild cattle so carved that the back line of the foreground aurochs defines the visible parts of the other, a use of occluding perspective 20,000 years old. This piece, moreover, was not hidden among the dramatic shadows within any cave but "decorated the place where people slept, ate, and cooked."

The last chapter of the text, treating most of the splendid photographs of painted, engraved or carved pieces, is aimed at grasping symbolic meaning, the ideas that underlie the powerful art. Here are most of those enigmatic Venus figurines (another, the first with full provenance, was found a decade ago), elaborately carved designs, tallies and counts, and the striking reliefs and engravings of animals, often superposed and reworked as if to record season after season by the riverbank. The late André Leroi-Gourhan and some of his successors saw overarching recurrent patterns in the placement of animal representations on the walls of the great decorated caves. Perhaps, the investigators speculated, the system modeled a dual male/female cosmology, or some other symbolic construction. In the absence of enough data for sharp statistical and comparative tests, the reader is perhaps justi-

Bison that were modeled in clay, deep in a cave, 15,000 years ago fied in wondering whether such an ordering, at once subtle and familiar, may owe as much to the influence of Descartes as it does to the French thinkers of the Ice Age.

The first images that are appropriately called art are firmly dated to about 35,000 years ago. Those earlier Neanderthal artisans certainly used ocher and manganese pigments, but the skeptical remind us that ocher does repel vermin and preserve skins. We have recovered fewer than a dozen objects that bear markings from the entire span of the Mousterian, although in the eastern Mediterranean the same people did bury their ochered dead in ceremony. Art manifestly flowers from the end of the Mousterian until the last of the Magdalenian, 11,000 years ago. Yet the finds are strongly discontinuous both in space and in time. The subsequent period, the Azilian, left us numerous painted pebbles bearing combinations of marks that may be an abstract symbol system, but no figurative art.

Above all, it is the images here that speak. They tell of a world of expressive genius, of irresistible empathy next to truly enigmatic and alienating form—in short, of the world of human beings. White writes: "Understanding the meaning and behavioral context of Upper Paleolithic art represents one of the greatest challenges facing archaeologists today." The generosity of the lenders who made this collection available for a while in the New World and the energy and insight of the author, guest curator for the show, give this inexpensive, comprehensive and beautiful book a special value. It deserves, but did not gain, an index.

EVERYONE HERE SPOKE SIGN LANGUAGE: HEREDITARY DEAFNESS ON MARTHA'S VINEYARD, by Nora Ellen Groce. Harvard University Press ($17.50).

This brilliantly argued and lively book first examines a concrete problem that random nature coldly set human beings and then traces out an unexpected solution devised by human nurture. Dr. Groce is a medical anthropologist at Harvard; her information consists of the oral history she herself garnered from some 50 witnesses, almost all more than 75 years old, and the documents in print and in manuscript that cross-check and extend their firsthand accounts. Human genetic theory, ethnographic counterparts and a clear-eyed account of social attitudes are the analytic tools that form her brief and telling work.

The famous Vineyard is 100 square miles of moor, woodland, field, pond and beach five salt miles off the Massachusetts mainland. Thomas Mayhew, Sr., of Watertown near Boston bought the rights to the island (and several other islands nearby) for £40 and a beaver hat in the year 1641. No Europeans were yet there at all, and the Indian bands assumed and remain in a reasonably peaceable stance toward their new neighbors. For a generation or two a few families moved to the island each year, first from lands near Boston and later from the nearby Cape Cod shore. By 1710 immigration had stopped: there were open lands elsewhere, whereas the island was settled.

After American independence Martha's Vineyard became a small maritime power, like Nantucket a nursery for the far-flung New England whalemen. But its westernmost township, Chilmark, a quiet up-island settlement with only a small harbor, lived by its subsistence farms, the sheep pastured among long stone walls, and one-man fishing dories, daring daily the pounding surf off South Beach. Few men there signed on as whalers, and even a day's trip to the hustle of the down-island seaports was a major event.

People lived close among their kin, in farm neighborhoods that could be traced back to the first settlers. In 1850 there were some 650 people in Chilmark, four-tenths of whom bore one of five surnames: Mayhew, Tilton, West, Hillman or Allen. The place was a genetic isolate, its lines of descent intricately knotted. New blood hardly entered the inbred up-island community. By the 1850's the percentage of those who could count no ancestor with a Kentish surname such as Tilton, one of the families that had come before 1710, was negligible.

The founders left to their people more than a name. In the mid-19th century the number of deaf individuals in Chilmark was about one in 25; 39 individuals are known to have been born deaf there in the course of two centuries, in a town whose average population was only 350. That rate is crudely 100 times higher than the rate for the entire American population in the 19th century.

Deaf parents did not always have deaf children, and hearing parents sometimes did. The origin of the condition was then a mystery. Not until Mendel's concept of recessive inheritance became known early in this century did the cause become clear. A mutation can alter the normal development of the complex circuitry of hearing, but the single mutant gene does not affect the one who carries it. It is passed at random, still hidden, to one of his or her children. (There is no sign of sex linkage, by the way; men and women were affected alike.) A great many genes induce recessive deafness; it is highly unlikely that a child can inherit deafness unless the parents share a common ancestry and each happens to donate the same hidden mutant gene.

Congenital deafness is gone from the island now. The last person to be born deaf there died in 1952. Chilmark today is no isolate. Young people go off-island to work and often return with a spouse. Many immigrants have married islanders. The American Asylum for the Deaf in Hartford by 1860 was enrolling many of the deaf children from the Vineyard. They were the largest group in the school. Quite a few married a deaf classmate. But there was much less chance of finding a carrier of the island recessive gene among that group than there was among hearing islander relatives.

Where did the mutant gene arise? The Vineyard families among whom deaf members were first documented knew each other well, but they were not related. They had all settled in Scituate south of Boston, the second-oldest town in Plymouth colony. Together they had moved to the Cape shore, and 25 years later on again to the Vineyard for new land. They had begun to intermarry; the birthrate was high and the death rate was low. They multiplied.

Those families were united by the sufferings of civil war, depression and dissent. Most of them had been Puritan clothworkers in the isolated poor, chalky lands of The Weald of Kent—once a wild forest, the largest in Roman England—40 miles southeast of London. The Reverend John Lothrop, their godly minister, led them to Boston in 1634, all migrating in a loyal party, most in a single ship. During the next years more "men of Kent" from nearby villages of The Weald joined them in Scituate. The Weald was no island, but its secluded villages were also something of a genetic isolate. When the Kentish folk settled at last on the Vineyard, among them were many carriers of that hidden mutant gene. Most of the deaf Vineyarders shared three early colonists as ancestors; probably all three of them had borne the gene.

In the spring of 1714 Judge Samuel Sewell of Boston engaged some island fishermen to guide his party to Edgartown after a long trip across Vineyard Sound. The judge was "ready to be offended that [one of the men] spake not a word to us. But it seems he is deaf and dumb." He was Jonathan Lambert, born in 1657 at Barnstable on Cape Cod of Kentish parents, the first recorded case of deafness on the island. There is no record of widespread deafness in The Weald of Kent. There is a hint, however, in the witness of Samuel Pepys himself. He noted during the Great Fire of London that a political acquaintance of his used sign language with deaf people engaged in intelligence work. That London official had grown up in a market town in The Weald around 1630.

Incomplete, the case is still compelling. Vineyard deafness was certainly the outcome of an old mutation. The place of that genetic event is almost surely within a dozen miles of the village of Egerton in The Weald of Kent, whence Lothrop and his congregants had come. The time remains uncertain; the single error could have occurred generations before the transatlantic migration.

The islanders' marriages with their neighbors shuffled the replicated gene widely enough so that the gene could often find its partner, and deafness, without other symptoms, could appear. Full expression came in the 1850's, after the gene had replicated for a dozen generations more or less in hiding.

Is molecular biology, to be sure with a good deal of help from society, to be viewed as destiny after all? Some people, plainly, were destined by nature to live without hearing. But chromosomes do not and cannot make a handicap out of that mere disability; it is society that handicaps. It was nurture, not nature, that fixed quite a different destiny for the deaf men, women and children who inhabited this engaging island.

Never were they treated as handicapped; never were they outcast, impoverished or isolated. The islanders instead actively adapted their society to the pattern of frequent hereditary deafness. "'Oh, they didn't think anything about them, they were just like everyone else...everyone here spoke sign language.'" For almost three centuries the deaf were part of the whole, free to vote, free to marry anyone, active in church affairs, legally fully responsible. They were farmers, militiamen, dorymen, carpenters. They were on hand at every social affair, all those merry chowder suppers, now winning and again losing at whist and checkers, practical jokers, once in a while celebrated horse traders or even memorable storytellers "in deaf and dumb," always part of the main. "As far as can be ascertained, deaf Islanders did not perceive themselves as a distinct social group."

Deafness just happened to some infants; it might enter any family as a minor problem of no deep importance, and it was best to get on with life. Nathaniel Mann, fisherman and dairy farmer, was born deaf; in 1924 he died the richest man in Chilmark. Tax and bank records indicate that there was in general no financial difference between deaf and hearing people. In today's America the deaf earn on the average about a third less than what hearing people earn.

Hearing and deaf islanders learned sign language in childhood, as naturally as the hearing learned spoken English. In hearing island families the children had no formal teaching in sign, but they picked it up from wide exposure; deaf adults and deaf children were important all around them. One old man reported that his hearing cousin learned to sign with her deaf father before she could talk with her hearing mother. Deaf children today are observed to gain vocabularies in sign as early as and at about the same rate as hearing children learn words; bilingual learning appears to be easy—sign and speech. Language is wider than speech, social communication is wider than language and our children are adept at it all.

Signers who could hear recalled that the syntax and word order of island sign was not the same as that of English; off-island, many hearing signers did follow English word by word rather than use the swifter forms of American Sign Language. Down-islanders who wandered into the Chilmark general store were often disconcerted by the silent general conversation passing around them among the locals, both the hearing and the deaf. The *Boston Sunday Herald* reported charmingly in 1895 on "singular pantomimes": neighborly morning gossip in sign carried out through spyglasses between Chilmark back doors an eighth of a mile apart.

It is now plain that sign languages are not dependent on spoken tongues. Most children born deaf have hearing parents who need to develop a home system of communication. Whenever there are enough deaf people in the course of time to pass along a system, a mature sign language may develop. There are recent studies of the use of signs by groups of deaf individuals in half a dozen villages around the world—in Yucatán, Surinam, Ghana, the Solomons....

The origin of signed languages was long held to be Enlightenment Paris, where the good Abbé de l'Épée invented it at his schools for the deaf, said to have been the first such institutions. The Abbé himself did not claim to have invented signing. He knew there was at least a regional sign language already in use among his pupils. It seems plain that such a sign language has served Vineyarders since Jonathan Lambert's time. It is likely to have come from Kent. Indeed, British sign language is little studied, although there is evidence of its use since the Middle Ages. It is probable that there were many regional dialects.

Once the deaf children from the Vineyard went off to school in Hartford at state expense, the Chilmark sign language began to merge with the official American Sign Language that had come from France. Groce's elderly hearing informants found it hard to understand deaf off-islanders, or the signers they saw on television; the Vineyarders must have signed in an interesting creole.

"I was born in Chilmark," writes John W. M. Whiting, a distinguished Harvard psychologist, in an eloquent foreword, "and am related through my mother to most of the families in that community.... I never learned to sign, but I saw sign language used...[at] the annual county fair...and when I went fishing.... Knowing how to use sign language was, like knowing French, something to be envied. With the influx of new residents...the incidence of hereditary deafness...declined and finally ceased altogether. But this benefit was countered by the destruction of a microcommunity as a strongly supportive network, something that is sadly missing in our modern industrial world."

We need to weave anew; this persuasive and compassionate investigation has uncovered a gleaming old design.

WAS EINSTEIN RIGHT? PUTTING GENERAL RELATIVITY TO THE TEST, by Clifford M. Will. Basic Books, Inc., Publishers ($18.95).

Was Einstein right? By the summer of 1974 a good number of pulsars were known. It was worthwhile to extend the list to catch fast ones. That might be done using the big radio dish at Arecibo, once you had a systematic means of searching the sky. Joseph H. Taylor and Russell A. Hulse of the University of Massachusetts had made a suitable computer program. Each day it would examine signals from a small strip of sky the dish scans as the earth turns. At the close of the day any recorded pulsing sources would be tested for well-defined rhythm. All the known pulsars were excellent clocks, most of them beating very steadily in the near-second range.

Graduate student Hulse worked day by day at the dish; Taylor, a young faculty member, stayed in Amherst and flew regularly to Puerto Rico to see how things were going. In early July a weak, repetitive signal caught Hulse's attention. It was a fast pulsar, if it was one at all; a pulse every six milliseconds meant that only the famous Crab spun more rapidly. That made the object well worth a second look; pulsars as fast as that are still uncommon. In late August, Hulse took that second look, this time intending to check the period carefully.

Genuine pulsars keep time to a microsecond a day. This object failed wretchedly: it slowed by tens of microseconds during a two-hour run. Maybe the ragged pulses were simply too noise-beset to measure well. "If Hulse had actually...dumped the candidate, he and Taylor would have been the astronomical goats of the decade." Instead Hulse kept at it, to find that the new program he hastily wrote to handle such a faint source still gave shifting results; from time to time the beat even sped up. At once he saw this was not all artifact. The shifts of one day were virtually repeated the next. Those data from Hulse's notebook are graphed here. Over a few days (the source was in the antenna beam only an hour or so each day) he could watch the pulse rate systematically slack off and speed up again, by tens of microseconds in a matter of an hour's run.

Now Hulse understood. The spinning pulsar-clock must itself be part of a double star, circling in orbit about an invisible companion. Orbital motion meant that the rhythm heard on the earth was Doppler-shifted, beating faster when the source was heading toward the earth, slower when it was moving away. This was the first binary pulsar ever seen. It was high time to telephone Taylor.

The binary pulsar nourished a cottage industry, generating a paper a week for two months and a great many more over the years. By now there has been a decade of detailed study, mainly led by Taylor. The pulsar that had appeared to be so inconstant is in fact one of the best clocks ever seen: its intrinsic pulse rate changes by only a fraction of a nanosecond in a year. All the rest of the complex rate changes observed reflect in detail that clock's orbital motion about its compact star companion (some small part is added by the earth's own motions). A circuit roughly the size of the circumference of our sun is traversed by the pulsar every eight hours.

Sophisticated analyses of these superbly precise measurements have by now mapped in complete detail the changing orbit of that bound pair of stars. Recall that there is a plethora of data: the huge sample of repeating pulses, half a billion a year, guarantees unprecedented accuracy. The pulsar is in fact a rather better timepiece than an atomic clock.

A feel for the precision of this kind of work may grow from the mere look of the data: the intrinsic pulsar period as of September 1, 1974, was just .059029995271 second. Certainly astronomers enjoy fine clocks; that is not the nub of the story. Rather, out of the security of timing that can be improved and rechecked year after year the binary pulsar has provided the most stringent tests we have of Einstein's general theory of relativity. It is a relativity laboratory far better than our own solar system.

The first test of the theory was recognized by Einstein himself, a small but inexplicable old discrepancy from the best Newtonian calculations. The orbital ellipse of Mercury slowly shifts its orientation at a rate predicted by general relativity. The binary pulsar displays that very same motion, but now so exaggerated as to be almost gross. The angle turned by the planet's orbit amounts to a couple of parts per million out of a circle during an entire year; the test with Mercury is therefore quantitatively still insecure. In the case of the pulsar, we find a robust and unmistakable annual four degrees.

The pulsar motions are now so fully determined by the richness of all these measurements that a pair of neutron stars 16,000 light-years away and hidden by dust are much better surveyed than any other star orbit anywhere. There is a final triumph for this signaling cosmic clock, so suited to checking by the laboratory clocks that are the best measuring instruments in the contemporary world. Year after year the binary-pulsar orbit is observed to shrink, although by only a very little. Those two stars as they orbit are gradually falling together. Where does that orbital energy go? There is no sign of any frictional effects that are large enough to matter.

The author explains the finding in the simplest of terms. Newton's action at a distance, the elegant inverse-square attraction, ignores time. Einstein's theory does not. For him a gravitational field takes time to propagate through free space, the same time light would take (provided the field is not too strong). This implies that the force felt by each star from the other lags a little behind the force expected from the current position of the partner. The resulting tiny mismatch of force direction means some of the orbital energy is lost from the moving masses into gravitational waves that carry it far away. A similar lag among currents flowing in the different parts of a radio-transmission antenna leads to the familiar loss of electromagnetic energy that ends up as the radio signal projected into space.

Gravitational radiation is extremely weak, however. We could never find the energy sent to us by such a distant, slow and stately system as two stars, even in tight orbit. Maybe some formidable sudden collapse can give us radiant gravitational energy enough to detect one day. But even without detecting the radiation we can reckon the energy lost on the basis of the unmistakable orbital changes as the two radiating stars circle closer.

Taylor first reported roughly the effect expected, just in time for the centenary of Einstein's birth in 1979. By now, after more than a decade of scrutinizing the binary pulsar, he and his colleagues have a refined result. Where the theory predicts a decrease in orbital period of 75 microseconds after a year of about 1,000 star orbits, the data show a decrease of 76 plus or minus two microseconds a year. Over the years slight changes in the course of the change themselves fit the predictions, which depend on the measured star masses. This work is a signal tour de force of experimental mastery and theoretical grasp.

Clifford Will is a physicist at Washington University who has been one of the most helpful of the theorists engaged in asking whether Einstein was right. He has written this personal and attractive account, with many instructive thought experiments and not a single formula, to draw the general reader into the exciting chronicle. For Will the 25 years that have elapsed since he was a 10th grader "convinced that genetics was [his] calling" have become the years of the renaissance of general relativity. Quasars and their energy puzzle, new workers and new ideas within the theory itself, the powerful new tools of planetary radar, lasers, nuclear resonance and space probes—even a stimulating theory or two offered as competitors to Einstein—all amounted to a new beginning for a taxing discipline turned formal and inbred for want of novelty.

The time delay in the motion of light as it passed the sun was proposed and found—an ingenious and brand-new test that became possible only when we could generate signals from beyond the sun—while planetary radar work was getting under way in the 1960's. After the Martian landers and their orbital radio relays could be used, the errors dwindled, until now the results agree with Einstein to about one part in 1,000, the most accurate test of the theory performed so far.

There are less orthodox proposals that challenge Einstein. One of these is the provocative idea that large masses and small ones would not fall in exactly the same way. This is no return to

Aristotle, but an effect that would indicate—some theories predict it—that the gravitational self-energy of any mass differs subtly from other forms of energy. Fifteen hundred runs of laser bounces from retroreflectors set on the lunar surface left no room for the new effect; the moon and the earth fall as one toward the sun. Those fixes are good to within about six inches out of the distance between the earth and the moon.

The sophisticated computing of the motions of the entire well-measured solar system of planets and satellites was input for another experiment: an effort to see whether the constant of gravitation changed over time, perhaps somehow in response to the expansion of the universe. The major uncertainty lies now in the unknown mass of all the asteroids that wrinkle the orbit of Mars and the other planets. Viking data limit the change in gravity to less than one part in 100 billion per year, but a naive guess would surely suggest that the value marches with the age of the universe, say one part in 20 billion per year. The leeway is too small; another factor of 10 is badly wanted. Perhaps a Mercury orbiter could be radio-ranged for a couple of years. That might reveal the effect, or limit it below a fraction of one part in a trillion per year. No such mission is planned.

Cliff Will has made his book as much a fresh and helpful reader's guide to the meaning of relativistic physics as it is the personal survey of the exciting state of theory testing here reviewed. The instructive part is in itself a useful book. Take only one instance of many: a triangle of ideal rulers makes pretty clear the bending of light by the curved nature of space.

As for the tests themselves, the author says this: "Would Einstein be pleased with the results? I think he would, since his theory has passed every one of the tests with flying colors. Would he be impressed? I'm not so sure.... He was apparently completely convinced of the validity of general relativity.... For better or worse, we of the late twentieth century have become more cynical."

A reviewer dares even harder words. None of these tests pushes beyond the domain of the cosmos where we dwell, far from really large compact masses, in weak gravity fields extending over small cosmic distances. The theory is surely not complete, for it is not yet a quantum theory, and in our day we take that ill. Albert Einstein was as right as a physicist can be; his theory will always be one cornerstone of our understanding of gravity. That it is final, that he was RIGHT, remains very hard to believe.

A meeting of two primates in the rain-forest canopy

THE ENCHANTED CANOPY: A JOURNEY OF DISCOVERY TO THE LAST UNEXPLORED FRONTIER, THE ROOF OF THE WORLD'S RAINFORESTS, by Andrew W. Mitchell. Macmillan Publishing Company ($29.95). **LIFE ABOVE THE JUNGLE FLOOR**, by Donald Perry. Simon & Schuster, Inc. ($16.95).

We know life on two-dimensional surfaces, usually on land, rarely at sea. The daring have become reasonably at home even within the waters, where the sunlight that throws life's shuttle has generated an intricate living web below, full of wonder and beauty. We are more intrusive in the air; our mastery of flight, for instance, is not such that our aircraft can accompany the birds as easily as the scuba diver joins a school of dolphin or colorful reef fishes.

These two attractive and readable books tell of new excursions into "yet another continent of life" that "remains to be discovered, not upon the earth, but one or two hundred feet above it.... There awaits a rich harvest for the naturalist who overcomes the obstacles—gravitation, ants, thorns, rotten trunks—and mounts to the summits of the jungle trees." It has been 70 years since the peerless naturalist-writer William Beebe wrote that passage, cited in both books.

The rain forests of the earth are regularly studied, but once scrutiny from the ground up has hit that green ceiling our knowledge dwindles sharply. It is just above the ceiling that the sunlight generates all the life of the forest; we poke among the fallen debris below and its denizens, examining the most intricate of all living patterns very partially indeed. It was not so for our ancestors; one marvelous photograph evokes that past. The Mitchell book shows "a meeting of two primates in the forest canopy," one a wary orang clinging gracefully high on a tall trunk, one a filmmaker dangling just as high by his steel cable, just two trees away.

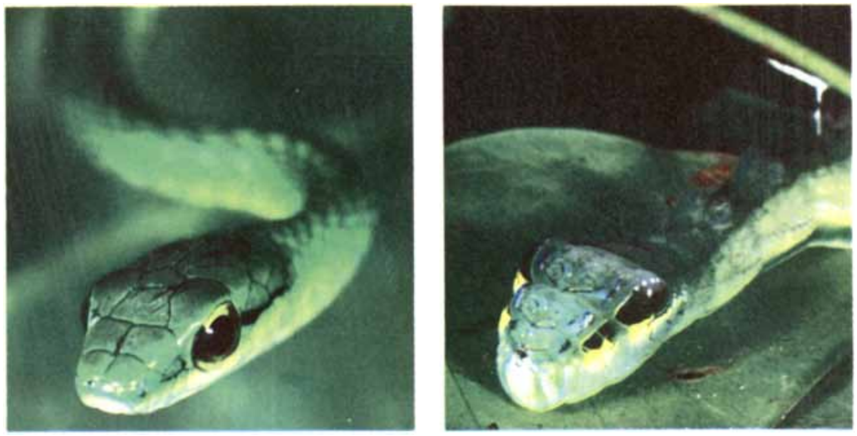
Entry into the canopy is becoming more usual. The harvest has already been rich, although the season is only beginning. A dozen years ago a naturalist raised high into the treetops a device that emitted a "fog of insecticide." So many unknown insects fell into his collecting trays below that he increased the estimate of the total number of insect species from one million to 30 million. Indeed, half of all animal species may be residents of that upper canopy.

Plant forms abound there too, epiphytes firmly rooted high above the forest floor. The leafy chalices of the pineapple kin, a couple of tons of watery pools commonly found on any big tree, are so many tiny wetlands, where even earthworms burrow the "aerial mulch," a soil fertilized by airborne particles, by falling leaves and by the unlucky little animals that live in those waters. There can be a frightening density of mosquitoes in the wet canopy, on occasion as high as 1,000 times the value measured on the floor. First found in the temperate rain forest of the Olympic Peninsula, a canopy root system now seems to be generally present among rain-forest trees. Wherever epiphytes abound, the trees that host them grow a complete root system out of their branches—roots that eagerly seek the nutrients held in that second soil overhead.

The canopy is to be seen at landscape scale from hills or from aircraft. It usually appears enigmatically bland, green, studded randomly with dying trees. The forest understory too usually disappoints the traveler, who expects tropical blossoms, with its endless, unrelieved green. But high above the dim floor "giant trees conduct their sex lives concealed within the privacy of the canopy." There is a fine photograph of rain-forest trees under the sun decked in crimson and yellow flower. The details of the flowering cycles are mainly a mystery.

Every half-dozen years in the rain forests of Malaysia and Indonesia many species of trees, all belonging to one genus, burst into near-synchronous blossom. Those trees flower over hundreds of square kilometers. At the Forestry Research Institute of Kepong investigators determined to study the event, unique to that region. By clever rigging of an aluminum spar held by block and tackle to a high branch, a bosun's chair was sent high aloft bearing an intrepid botanist.

Contrary to expectation, the flowers were not meant for wind pollination. They opened fragrant with nectar only at night. Night duty was scary but revealing: the flowers were full of thrips, two-millimeter insects with feathery wings, the pollinators of the towering forest. The insects did not leave the flowers where they fed. Every morning the blossoms fell 100 feet and more to the forest floor, a couple of hundred thousand blossoms from each tree each day, a few thrips falling within every blossom. At dusk the thrips flew by the million back up to fragrant new flowers, carrying with them the sticky pollen. Because the insects are so small, even the gentle air currents of the forest can divert them to neighboring trees for cross-fertilization.



Mimicry in the canopy: a parrot snake **(left)** *and its caterpillar double* **(right)**

Dr. Perry is an American biologist now engaged in building a specialized canopy-research system, able to reach a large volume of rain forest, including altitudes well above the treetops [see "The Canopy of the Tropical Rain Forest," by Donald R. Perry; SCIENTIFIC AMERICAN, November, 1984]. He and his engineering colleague, John Williams, have designed a kind of radio-controlled chair lift. Its stainless-steel cables will be anchored in high rock ridges above a small, isolated valley in the virgin rain forest in the foothills of northern Costa Rica, near the forest research station of Finca La Selva. Two investigators will be able to place themselves anywhere over a 10-acre area of forest roof and move down to the height they choose, in complete safety. Small platforms might be built to hang below the cable web as well, fitted out to support extended stays in the canopy. Perry's book is the shorter and more personal of the two, centered on his Costa Rican experience, but with free and interesting excursions into evolutionary reflections.

Andrew Mitchell is an English naturalist and television producer. He has ventured into floor and canopy around the world, from La Selva to Kenya, Papua, Borneo. His book is personal and nontechnical, but longer and more comprehensive in its survey of canopy biology. He includes some history of the way into the canopy. It was the Oxford expedition to British Guiana in the late 1920's that first tried to enter the canopy systematically. Of course many travelers had now and again climbed a tree of the rain forest; the first English account of such an effort was that of Sir Francis Drake, who once glimpsed both oceans from a high forest tree in Darien, a vista that inspired his audacious incursion into "that far-off forbidden sea."

Mitchell summarizes the efforts to gain cheap, safe, flexible entry into the canopy. There are aerial walkways, resembling light suspension bridges, that run from tree to tree; climbing ropes using the clamps and ascenders of the mountaineers and cavers, scaffolding towers and Perry's automated web. Helicopters have been tried; they are costly, and the downdraft is disturbing to much life of the canopy. A hot-air balloon lofting a kind of aerial raft, a light inflated platform as big as a tennis court, is under test in French Guiana. A camera-carrying hydrogen balloon tethered to a truck on a forest road has been used in Gabon. Ultralight aircraft have been tried for wider survey. There are dreams of small two-person blimps with quiet engines, and perhaps of still smaller tethered ones bearing video cameras and sampling devices, rather like the remote-viewing vehicles that have been so successful on the sea floor.

Beebe was right: the rewards to the arboreal naturalist are rich. Among them may be the means and motives for preserving the rain forests themselves, in particular against the shortsighted actions of our own species, ungrateful emigrants that we are from a world our forebears knew well before ever they walked on solid ground.

Established 1845 SCIENTIFIC AMERICAN March 1987 Volume 256 Number 3

Dyslexia

Mirror writing and similar problems are usually blamed on defects in visual perception, but in truth dyslexia seems to be a complex linguistic deficiency. The remedy is proper instruction in reading

by Frank R. Vellutino

Dyslexia is a generic term that has come to refer to an extraordinary difficulty experienced by otherwise normal children in learning to identify printed words, presumably as the result of constitutional deficiencies. The condition is commonly believed to originate in the visual-spatial system. Its presence is considered to be signaled by mirror writing and letter reversal. Dyslexics, it is believed, show uncertain hand preference. Children whose first language is based on alphabetic rather than pictographic or ideographic characters are said to be particularly susceptible to the condition. Finally, dyslexia is widely considered to be correctable by means of therapies aimed at "strengthening" the visual-spatial system. Each of these perceptions, contemporary research shows, is seriously flawed.

It was through the work of the U.S. neuropsychiatrist Samuel Torrey Orton in 1925 that the deficiency first came to be perceived as lying in the visual system. Orton suggested that an apparent dysfunction in visual perception and visual memory, characterized by a tendency to perceive letters and words in reverse (*b* for *d* or *was* for *saw*), causes dyslexia. Such a disorder would also explain mirror writing. Orton further suggested that the disorder is caused by a maturational lag: the consequence of a failure of one or the other hemisphere of the brain to dominate the development of language. This last proposal, at least, was and is still a viable hypothesis.

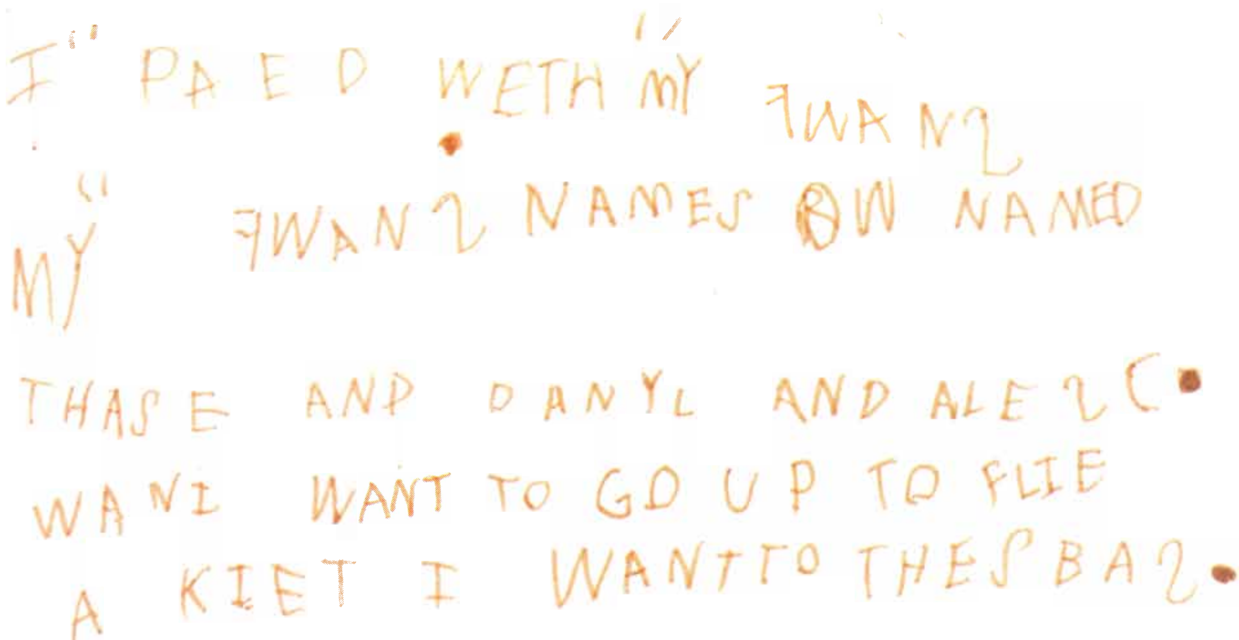
Related hypotheses, not attributable to Orton, are that dyslexia may somehow be caused either by motor and visual defects or by eye-movement defects affecting binocular coordination, eye tracking and directional scanning. Both the concept of dyslexia as a visual problem and its presumed association with uncertain cerebral dominance still underlie many therapeutic approaches to the condition.

Working at the Child Research and Study Center of the State University of New York at Albany, my colleagues and I have begun to examine, and to challenge, common beliefs about dyslexia, including the notion that the condition stems primarily from visual deficits. Along with other researchers in this country and abroad, we have been finding that dyslexia is a subtle language deficiency. The deficiency has its roots in other areas: phonological-coding deficits (inability to represent and access the sound of a word in order to help remember the word); deficient phonemic segmentation (inability to break words into component sounds); poor vocabulary development, and trouble discriminating grammatical and syntactic differences among words and sentences. Far from being a visual problem, dyslexia appears to be the consequence of limited facility in using language to code other types of information.

To understand this definition of the disorder, one can conceive of the mind as an extremely sophisticated reference library. The library model is appropriate because recent studies seem to indicate that dyslexia is as closely aligned with the cross-referencing and retrieval of coded information already stored in memory as it is with the storing and coding of new information.

The library model is based on the assumption that the processing of information to be stored in memory proceeds in stages. The first stage of processing takes place in a sensory storage system, where a replica of a given stimulus is held briefly. The second stage is believed to take place in a short-term "working" memory: a limited-capacity system in which an encoded (transformed) version of the stimulus is available for no longer than 30 seconds. In this working memory, physical information is transformed into a more abstract symbolic representation for storage in long-term memory, which is thought to have an unlimited capacity. During the final stage of memory processing, the encoded form of the stimulus is either categorized and stored in long-term memory, discarded or inadvertently lost from working memory.

In research based on this model we have found that dyslexia is more a symptom of dysfunction during storage and retrieval of linguistic information than it is a consequence of a defect in the visual system. In one experiment poor readers in the second through sixth grades, who frequently made reversal errors, were asked to copy designs, words, scrambled letters and numerals after brief visual presentation. Afterward they were asked to name the stimuli that were actual words. We found that the poor readers could reproduce the letters in a stimulus word in the correct orientation and sequence even when they could not name the word accurately. For example, they typically copied *was* correctly but then often called it "saw." When asked to read out the letters of the words right



PRESUMED SIGNS OF DYSLEXIA such as mirror writing are often seen in the early stages of normal development of writing skills. At three Amanda mirror-writes her name (*top*). The habit persists when, at four, she writes about "my blanket I love the best" (*middle*). At five she writes about playing with her friends Tracy, Daniel and Alex. Mirror writing has all but disappeared; there is a clear ability to encode a desired sound by means of phonologically appropriate (if not always correct) letters (*bottom*).

after naming the words as wholes, they could name the letters of most words in the correct order even when they named the words incorrectly.

The inference from this experiment was clear. Errors such as calling *was* "saw" are the result of difficulties in storing and retrieving the names of printed words rather than of a dysfunction in visual-spatial processing.

This inference was reinforced by the results of a series of studies of children's ability to reproduce, from visual memory, words from an unfamiliar writing system. Groups of dyslexic and normal readers were asked to print Hebrew words and letters in the proper sequence and orientation after brief exposure. Some children who were already learning to read and write Hebrew were also tested to see how they would compare with the first two groups.

The important finding here was that the dyslexic readers did as well as the normal ones on this task—although, to be sure, neither group did as well as the children who were learning Hebrew. The result seemed to underscore the fact that when complex, wordlike symbols lacked any linguistic associates—had no meaning or sound, in effect—the visual recall of those symbols was no less difficult for the normal readers than it was for the poor readers. This implies that memory for visual symbols representing words is mediated by the linguistic properties of those words, particularly their meanings and sounds.

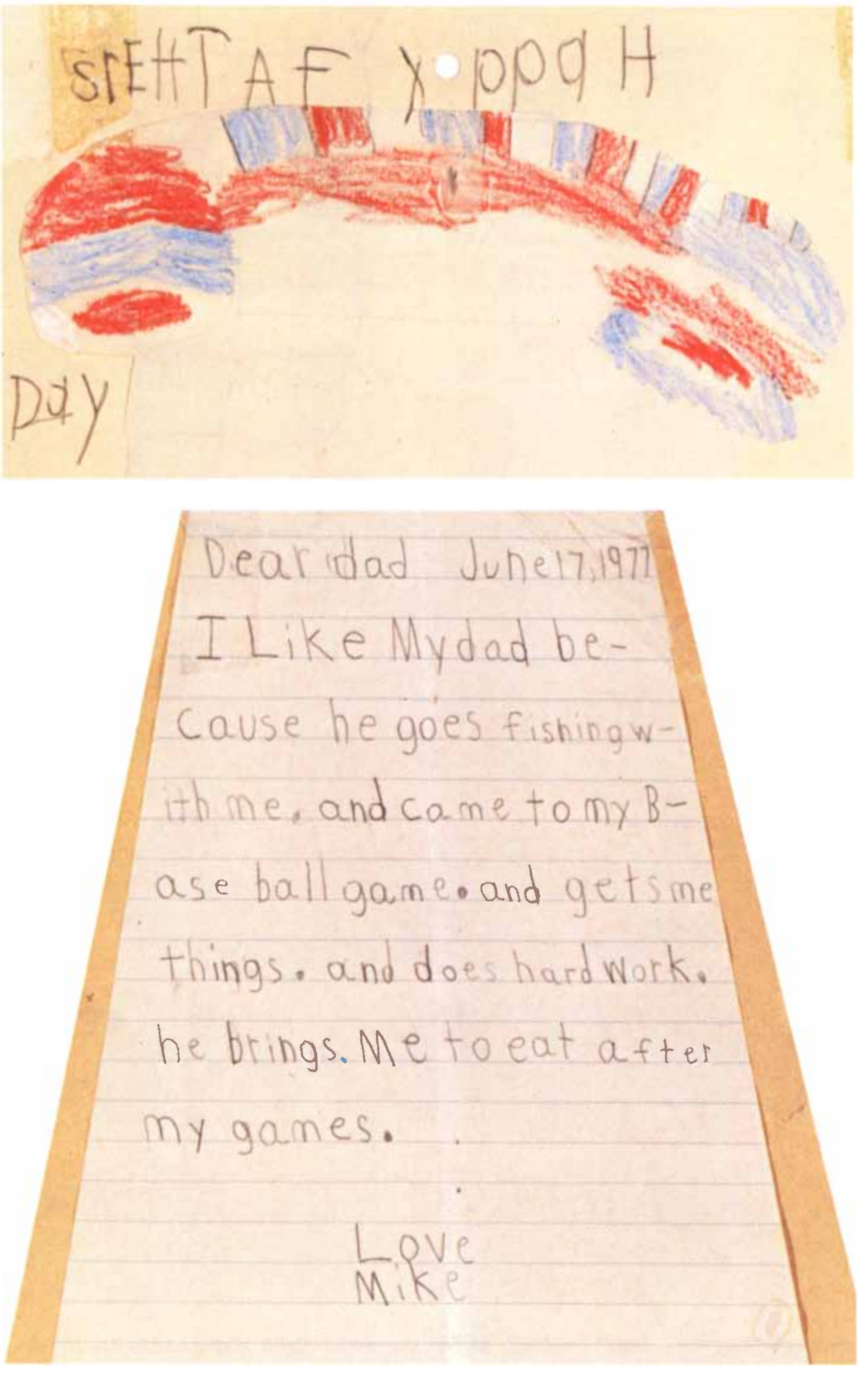
We found too that both groups of children who were not familiar with Hebrew manifested identical tendencies to process the Hebrew letters from left to right, suggesting that dyslexics are not inherently impaired in their ability to maintain left-to-right directionality. In other words, if dyslexics have difficulty maintaining proper directionality, it is a symptom of a reading disorder rather than the cause of the disorder.

A secondary inference from this study (one that has been documented by other studies directly evaluating sensory storage in dyslexic and normal readers) is that the dyslexic readers were able to hold a memory trace for as long as the normal readers. That is to say, visual traces dissipated no more rapidly in a dyslexic's sensory memory than in a normal reader's: visual form perception seems to be comparable in the two groups.

If dyslexic readers are at least capable of perceiving and reproducing letters at roughly the same level of accuracy as normally developing readers, then the problem is again thrown back on linguistic rather than visual coding systems. Printed words can be identified either through whole-word processing based on their salient visual features, their meanings and the context in which they appear, or through "part-whole" processing based on alphabetic mapping: breaking down words into letter sounds.

Because learning to read is inherently difficult, the beginning reader must be able to adopt both strategies to identify words. If the child leans too heavily on a whole-word strategy—and does not use the sounds associated with alphabetic characters to help decode new words—visual memory is inordinately taxed; errors such as *was/saw* and *lion/loin* result. On the other hand, children who rely exclusively on alphabetic mapping—and do not use salient visual features, word meanings and context to facilitate word identification—find it hard to read fluently and have trouble understanding what they read.

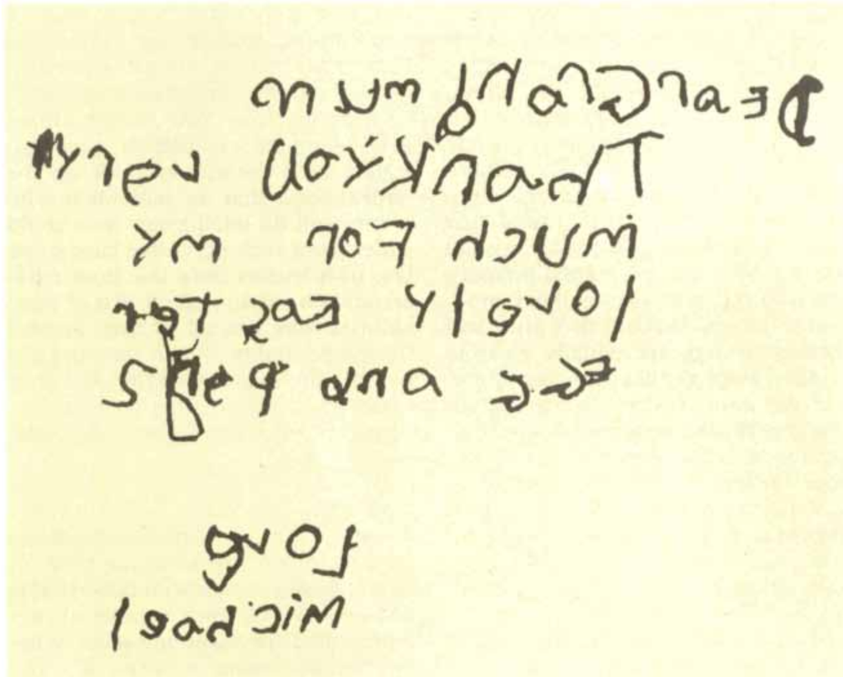
The implication is clear that dyslexia may be caused, on the one hand,



NORMAL DEVELOPMENT is evident in Father's Day cards written by a boy in first grade (*top*) and then in second grade, after he had a year of reading instruction (*bottom*).

by highly specific linguistic deficiencies (such as vocabulary weakness or sound-mapping deficits) affecting only certain of the subskills that are necessary components of the ability to read. On the other hand, it is equally plausible that more general language deficiencies are present, which affect all subskills. A hypothesis supporting the first view has been put forward by Isabelle Y. Liberman and Donald P. Shankweiler of the Haskins Laboratories, Inc., in New Haven.

According to Liberman and Shankweiler, poor readers are not explicitly aware that spoken and printed words can be segmented into individual phonemes; this makes it hard for them to learn to identify words through alphabetic mapping and letter-sound synthesis, or what is termed phonetic decoding. Poor phoneme segmentation is said to be a manifestation of a more general problem in phonological coding, characterized by the storage in memory of impoverished representations of letter sounds and word names.



TYPICAL MIRROR WRITING is displayed in a note from a five-year-old boy to "Grandmum." This child's development was not normal: when seen at 11, he was dyslexic, according to the English investigators Macdonald Critchley and Eileen A. Critchley.

Such a dysfunction could theoretically lead to difficulty not only in learning the sounds associated with given letters and combinations of letters but also in learning the names of printed words as whole entities. Words are therefore stored without complete phonological codes—file cards, in the library model. Asked to call up the proper word, the child finds that he or she has not retained enough clues to the name of the word.

Results of other studies done in our laboratory and elsewhere support the concept that deficiencies in alphabetic mapping and phonetic decoding are major factors in reading difficulties. The studies show consistently that severely impaired readers are much less proficient than normal readers in learning to use letter sounds to decode pseudowords (meaningless wordlike letter assemblages used in testing) and words they have never seen before. Such deficiencies seem to be the result of a poor grasp of phoneme values. It has also been found that kindergarten and first-grade children who have some ability to segment spoken words into syllables and phoneme-size units learn to read better than children who have little or no such ability. Perhaps the most impressive support comes from studies showing that children trained to identify phonemes have an increased ability to map alphabetically and therefore an enhanced capacity to identify printed words.

If relative lack of awareness of phonemes and deficiency in phonetic decoding are rooted in more basic difficulties in phonological coding, one might expect to find that poor readers have trouble remembering words they hear. This turns out to be the case. In a large number of studies done in our laboratory, at the Haskins Laboratories and elsewhere, poor readers did less well than normal readers when they were asked to recall lists of words they had just heard.

A number of investigators—notably Martha B. Denckla and the late Rita G. Rudel of the Columbia University College of Physicians and Surgeons—have also found that dyslexics tend to be slower and less accurate than normal readers not only in naming letters and words but also in naming common objects, colors and numerals. The performance of many dyslexics in these studies was often characterized by severe blocking: circumlocutions, long hesitations and such substitution errors as saying "dog" when confronted with a picture of a cat.

Deficiencies in vocabulary development and semantic ability in general also seem to make the identification of words difficult, as several studies have suggested. A deficiency in syntactic competence may be yet another factor. Studies in our laboratory and elsewhere show that poor readers seem to be less proficient than normal readers in comprehending sentences, particularly those that are syntactically complex; in making use of inflectional morphemes (such as *-ed* and *-ing*) to specify such things as tense and number; in distinguishing between grammatical and ungrammatical sentences; in using complex sentences in a grammatically correct way, and in making fine distinctions among abstract words, particularly such "noncontent" words as *if, but* and *their*.

In fact, considerable evidence suggests that poor readers are more deficient (compared with normally developing readers) in their ability to identify noncontent words than they are in their ability to identify content words such as *dog* or *cat*. Poor readers also seem to have difficulty using sentence context to help them identify printed words. It should be pointed out, however, that a causal connection between reading disability and deficiencies in processing the semantic and syntactic attributes of printed words has not yet been established.

The possibility that dyslexics may be impaired by deficiencies in language raises the question of whether or not they are basically impaired in auditory processing. One possibility is that their auditory sensory, or echoic, memory is deficient, which would mean the auditory trace is dissipated faster in poor readers than in normal readers. This possibility has been evaluated and dismissed by Randall W. Engle and his associates at the University of South Carolina. We have obtained similar results.

A second possibility is that poor

readers are generally limited in their ability to store acoustic information in permanent memory. Susan Brady of the Haskins Laboratories evaluated this question by comparing the ability of dyslexic readers and of normal readers to remember verbal and nonverbal information (words and environmental sounds). They found that dyslexics performed below the normal readers only on the verbal-memory tasks. In terms of our memory model such results indicate that poor and normal readers are equally capable in initial-stage auditory processing and that the poor readers do not sustain any generalized deficiencies in memory for material presented auditorily. Poor readers do, however, appear to be deficient in their ability to retrieve linguistic representations stored in long-term memory. Such results seem to be consistent with linguistic-coding theories of reading disability.

Several other hypotheses propose that the cause of dyslexia lies in nonlinguistic functions, but none of them is very persuasive. One, the "attention deficit" hypothesis, relates difficulty in reading to a generalized inability to concentrate and pay attention. Some workers have found evidence associating this pattern with physiological abnormalities. Children exhibiting it, however, have difficulty with subjects other than reading and may not be representative of those whose problems are limited to reading.

Another theory relates dyslexia to deficits in "cross-modal transfer," or an inability to relate stimuli perceived through one sensory system to stimuli perceived through another system. The theory suffers both from lack of experimental support and from logical consistency; it would seem improbable that a child whose intelligence is average or above average (as is true of dyslexics) could suffer from cross-modal-transfer deficits, given the degree of cross-modal learning required to score in at least the average range on any test of intelligence.

Some other theories deserve mention. One posits deficiencies in associative learning. Another argues that dyslexics have trouble detecting patterns and learning "invariant relationships," such as the rules that govern alphabetic mapping or number concepts. Again, both theories stumble on the unlikelihood that an individual who scores well on intelligence tests could suffer from such pervasive handicaps. Our own studies show that poor readers who do not do well on tests of these abilities were limited by their linguistic-coding ability. When the tests did not rely on linguistic coding, the poor readers were able to improve their performance on association and rule-learning tasks.

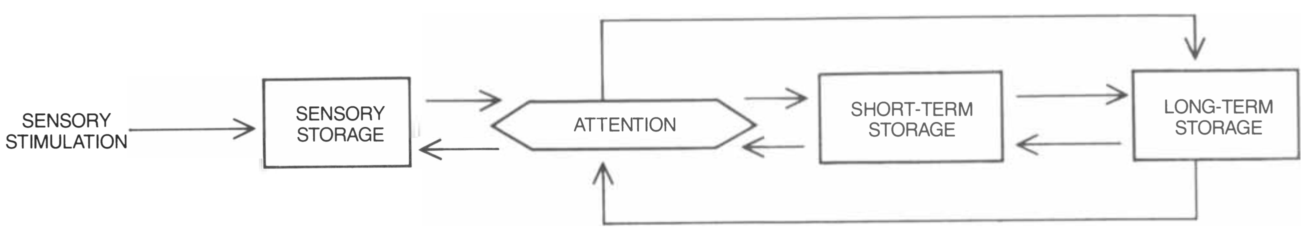
For example, in one such study we asked a group of poor and normal readers to learn to associate pairs of novel visual symbols with two-syllable nonsense words; each symbol always represented the same nonsense syllable [*see illustration on page 40*]. The child's task was to learn to say the two-syllable nonsense word when presented with the pair of symbols representing that word. To help remember the whole word the child was told to try to remember the individual symbol that represented a given syllable. After practice on this task the symbols were rearranged; the child was then shown the rearranged set of symbols to see if he or she could transfer, or generalize, the symbol-syllable units learned initially to associate the rearranged pairs. This is analogous to learning letter sounds in *cat, ran* and *fan* and relying on the sounds to decode "new" words such as *fat, rat* and *can*.

A second group of poor and normal readers was given a similar transfer-learning task, but instead of learning visual-verbal associates these subjects learned to associate and transfer visual pairs; again one of the symbols in a stimulus pair always represented one of the symbols in a response pair. One significant finding from this study was that poor readers did less well than normal readers on the visual-verbal learning tasks. The implication was that they had difficulty with both initial-association and transfer learning because they were impaired by inability to remember the nonsense syllables, not by inability to associate or generalize. This conclusion was verified by a second finding, which was that poor readers did as well as normal readers on the visual-visual association and transfer-learning tasks. Poor readers, then, appear to have difficulty with association and rule-learning tasks only when the tasks require them to store and retrieve the auditory representation of words and syllables.

I come, finally, to what might be called the serial-deficit theory. It holds that an inability to remember the sequence of a series of items or events underlies dyslexia. The theory presumes that the brain is equipped with a ubiquitous ordering ability, a supposition that strikes me as highly unlikely. I suspect instead that different cognitive systems have their own rules and algorithms for establishing order and sequence. If this is the case, then such theories are ruled out.

A number of investigators have taken an approach somewhat different from the ones discussed so far. They suggest that reading disability may be associated with a number of different neurologic disorders, each of which underlies one or another of the basic processes involved in learning to read. Such a view implies that there is an array of neurologic disorders characterized respectively by visual deficits, language deficits, deficiencies in cross-modal transfer and so forth. Although reading disability may result from a number of different factors, the cause in an otherwise normal child would seem to be more circumscribed. As I have argued, we believe the problem lies in the linguistic domain. The question remains open, however.

Our research and the work of investigators elsewhere has called into question other perceptions of dyslexia. First, it is important to point out



INFORMATION-PROCESSING MODEL of the stages of memory and their interrelations is diagrammed. A literal copy of a visual or auditory stimulus is held in sensory storage briefly. If the subject attends to the stimulus, it enters a short-term memory system, where it is encoded into a representation appropriate for storage in long-term memory; a stimulus that is not attended to is not encoded and is lost from memory. The long arrows suggest some of the interactions among the memory systems. For example, a person is more likely to attend to a familiar stimulus because a representation of it has been filed in long-term memory.

REAL WORDS		
THREE-LETTER	FOUR-LETTER	FIVE-LETTER
was	**loin**	**blunt**
SCRAMBLED LETTERS		
THREE-LETTER	FOUR-LETTER	FIVE-LETTER
dnv	**jpyc**	**ztbrc**
NUMBERS		
THREE-DIGIT	FOUR-DIGIT	FIVE-DIGIT
382	**4328**	**96842**

VERBAL AND NONVERBAL STIMULI were presented for half a second to poor and normal readers in the second and sixth grades; examples of each are given here. In one phase of the experiment subjects were asked to write down the words, scrambled letters or numbers from memory; in a second phase they were asked instead to name each character of a stimulus in the right order—in the case of the words, after pronouncing them. Poor readers did about as well as normal readers on the copying task but not on the naming; their problems seem to arise from deficiencies in verbal, not visual, processing.

that there are no well-defined reading behaviors that can clearly distinguish a dyslexic from other poor readers whose difficulties stem, for example, from limitations in experience; nor are there distinguishing clinical patterns. All poor readers have difficulty learning to identify and spell printed words, but not all would qualify as dyslexics—if the term dyslexia is used to define a very specific reading disability in an otherwise normal child. Moreover, the reversal errors said to be characteristic of dyslexics account for no more than between 20 and 25 percent of their reading errors, most of which are generalizations promoted by an imperfect knowledge of linguistic associates (such as *cat* for *fat, cat* for *kitty, bomber* for *bombardier*).

Moreover, reversal errors can be plausibly explained without invoking spatial confusion. If, for example, a child attempts to remember the words *pot* and *top* only as whole entities and does not know the sounds of the individual letters, there is a stronger inclination to reverse them. We have verified this in experiments in which two groups of dyslexic and normal readers (in the second and sixth grades) learned to identify pseudowords constructed from a novel alphabet, which were designed to prompt reversal errors of the *was/saw* type. Children taught to identify these pseudowords by a whole-word method made many more reversal errors than children who were taught to use alphabetic mapping, but the poor readers made no more reversal errors than the normal readers. It seems clear that the spatial-confusion interpretation of reversal errors is incorrect.

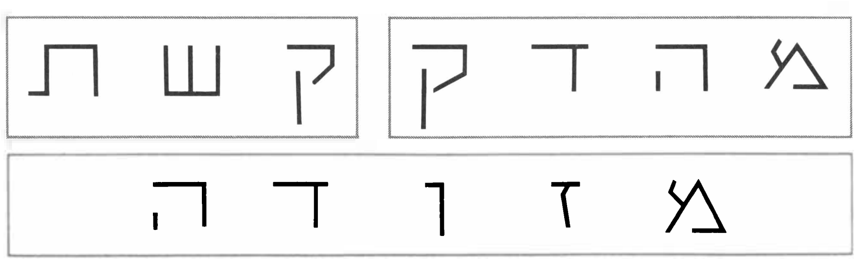
The clinical significance of mirror writing is also commonly misunderstood. Some degree of mirror writing can be observed in normally developing readers as well as in poor ones. The tendency is quite likely a vestige of an earlier stage of development, which some poor readers take more time to transcend. It persists in these children, I think, because they find it difficult to remember both the visual-linguistic clues and the clues provided by sentence context that foster accurate judgment about the relative positions of words and letters and the direction in which they run. What causes the habit of mirror writing to persist is a lack of correct practice in writing and spelling that actually results from a child's reading problems. I suggest, in other words, that when mirror writing is observed in poor readers, it is a consequence of their reading difficulty rather than a benchmark of visual-spatial confusion.

Another common misconception is that reading problems are caused by perceptual deficits associated with motor and visual-motor defects or with ocular defects other than loss of visual acuity. If deficiencies in motor and visual-motor development or defects in eye movements caused perceptual impairment and reading problems, one would be at a loss to explain how so many children with cerebral palsy and various visual-tracking defects become literate.

Still another misconception is that dyslexia is more prevalent in countries where the writing systems are based on an alphabet than it is in countries where the writing systems either are pictographic or are phonetically less complex. A study done by Harold W. Stevenson and his colleagues at the University of Michigan has yielded some evidence against this idea. The investigators evaluated schoolchildren in comparable cities in the U.S., Japan and Taiwan. The three groups were compared on a large battery of tests measuring school achievement as well as language and cognitive ability. The results yielded no evidence that the writing systems of Japan and China preclude reading disability. The fact that the Japanese and Chinese languages respectively contain characters representing syllables and entire words seemed to make no difference.

Since dyslexia appears to be commonly associated with brain dysfunction and the brain's ability to store and retrieve information, it is necessary to consider whether constitutional factors—genetic and/or neurologic—contribute to the condition. There are actually no diagnostic criteria that enable one to distinguish clearly between constitutionally and experientially derived origins of reading disability. There is one highly suggestive piece of evidence, however: boys who are impaired in reading outnumber girls who are so impaired by ratios ranging from 4 : 1 to 10 : 1.

Taken together with results from developmental studies showing that boys are in general less capable than girls on language and language-related tasks, such ratios could be taken as support for both constitutional and language-



HEBREW WORDS in simplified block form were presented one at a time to three groups (second through sixth grade): poor readers and normal readers unfamiliar with Hebrew and children currently studying it. Three-, four- and five-letter words (sampled here) were shown for three, four and five seconds respectively. Subjects had to reproduce the words on paper from memory. Among those unfamiliar with Hebrew, poor readers did as well as normal readers. (Neither, of course, did as well as children studying Hebrew.)

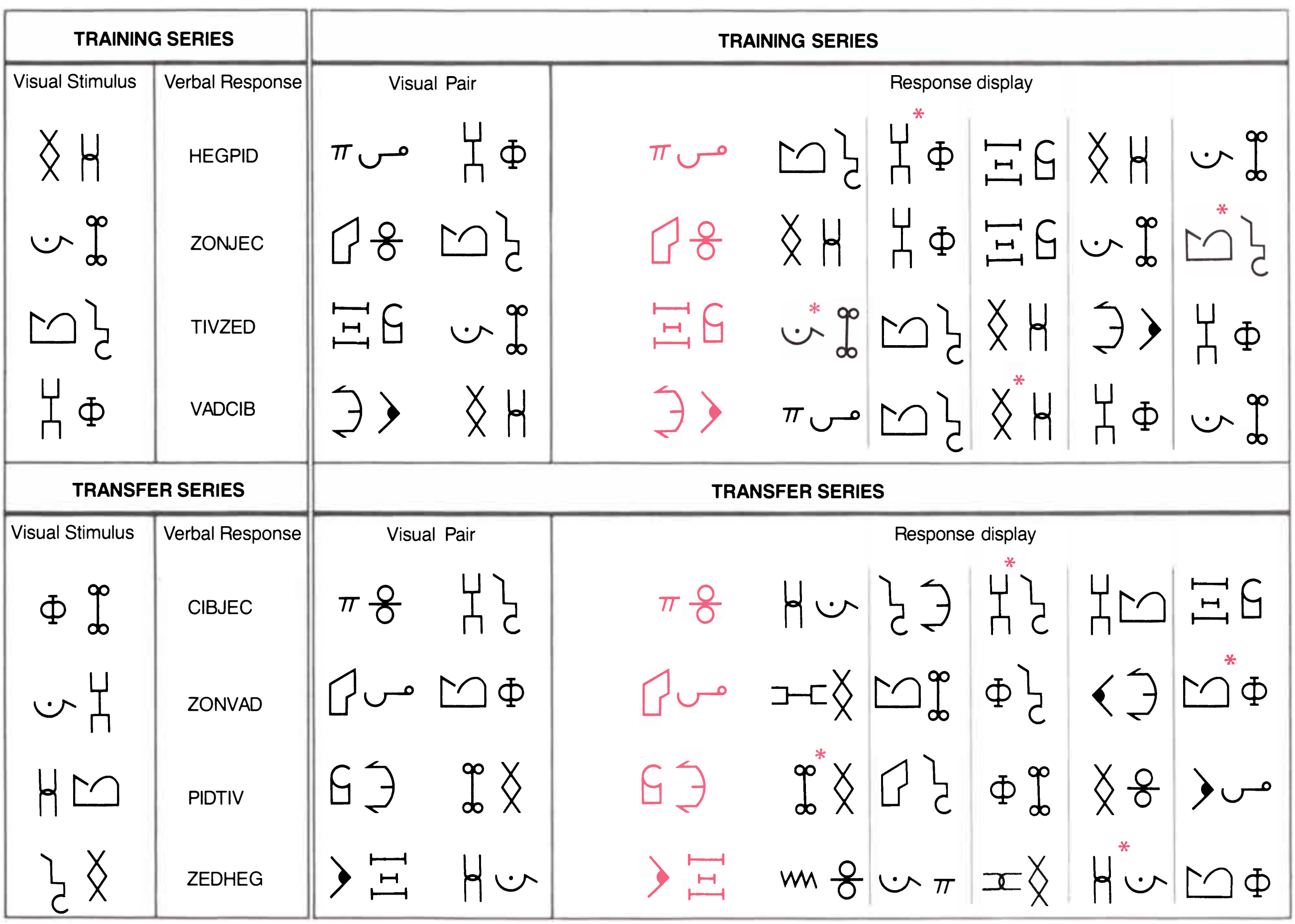
deficit theories of dyslexia. Boys may be either genetically less well endowed with linguistic capabilities than girls are or more vulnerable to neurologic defects affecting language development than girls are. Then, if it is true that reading disability is caused by limitations in language ability, constitutionally derived language deficits would result in a higher incidence of the disorder in boys than in girls.

Support for a genetic basis for dyslexia comes from a small number of familial and twin studies carried out over the years. The early studies in the literature were poorly controlled, but they consistently showed that reading disability occurs more often in near relatives than in the population at large, that reading disability occurs more often in twins than in siblings and that reading disability has a much higher concordance rate in monozygotic (identical) twins than in dizygotic twins. These findings have been verified recently in a more highly controlled study conducted by John C. DeFries and his associates at the Institute for Behavioral Genetics of the University of Colorado at Boulder.

Perhaps an even more exciting finding, from research done by the Boulder group with Shelly D. Smith, is the tentative localization of a particular gene on chromosome 15 in members of families in which there is a history of reading disability. Once a gene that may be responsible for a specific attribute has been localized on a specific chromosome, geneticists are in a position to find the mechanisms whereby the gene gives rise to the attribute. The finding could be a significant breakthrough in the study of dyslexia, but it has not yet been replicated.

A number of investigators of brain function have also begun to study the etiology of dyslexia, and their early findings are promising. A Boulder team headed by David W. Shucard compared dyslexic and normal readers on measures evaluating electrophysiological responses to auditory and visual stimulation. The group's major finding has been that electrical activity in response to reading is characterized in dyslexics by greater amplitude in the left hemisphere than in the right hemisphere, whereas the opposite pattern emerges in normal readers.

A novel technique recently exploited in this area of inquiry is "brain electrical activity mapping" (BEAM), developed by Frank H. Duffy and his colleagues at Children's Hospital in Boston. The BEAM technique produces topographic maps that are based on brain functions, not brain structures. Duffy and his associates have now obtained evidence that left-hemisphere functioning in dyslexics is qualitative-



RULE-LEARNING TASKS were given to poor and normal readers in the fourth, fifth and sixth grades. The visual-verbal task (*left*) involved novel visual symbols, each representing a syllable. Subjects were trained (*top*) to match each pair of symbols (stimuli) with the correct response, a two-syllable nonsense word. Then they were shown a transfer series (*bottom*), in which the symbols and words were reordered, and were asked to say the new words. In the visual-visual task (*right*) visual responses were substituted for the verbal ones. Now each stimulus-response pair was a pair of two-symbol "words." The response display presented the stimulus (*color*); the subject was trained to select the correct response from among a series of five pairs. The colored asterisks designate the correct responses. Again the symbols had been reordered in the transfer series, but the correspondence between individual stimulus and response symbols was maintained. The poor readers failed to perform as well as the normal readers only on the visual-verbal task, which suggests that they had difficulty learning rules only when the experiment required them to make a verbal response.

ly different from that in normally developing readers. The differences were particularly prominent in adjacent regions of the left parietal and temporal lobes, areas of the brain known to support speech, language and related linguistic activities.

Finally, I should mention the neuroanatomical studies conducted by Albert M. Galaburda and his colleagues at the Harvard Medical School. Two types of anatomic anomalies were disclosed by postmortem analysis of the brains of several male dyslexic subjects. First, there was a consistent absence of the standard pattern of brain asymmetry in language regions: atypically, the left-hemisphere regions were no more highly developed than the right-hemisphere regions. Second, in the cerebral cortex of the language-related areas there were multiple sites where the microarchitectural arrangement and position of neurons, or nerve cells, was distorted, particularly in the left hemisphere.

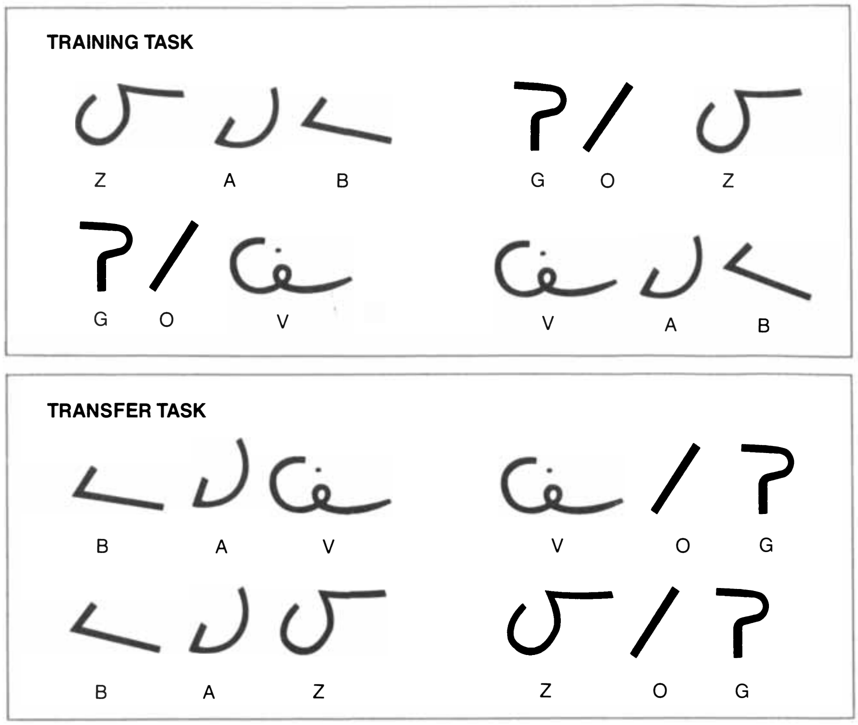
Since the absence of asymmetry resulted from the excessive development of right-hemisphere areas that are normally smaller, Galaburda has argued that both the anatomical anomaly and the anomalous neuronal architectures may reflect interference with the normal developmental process whereby undesirable neurons and their connections are eliminated. Presumably the oddity in microanatomical organization of the language areas of the brain would explain some linguistic deficits seen in the reading disorder.

The late Norman Geschwind hypothesized that the variations from the standard pattern of cerebral lateralization observed by Galaburda were significantly linked to disorders involving an immune dysfunction. He suggested that common developmental mechanisms acting during fetal life might lead to the abnormal development of parts of the immune system and also to anomalous asymmetrical development of the brain; he proposed an important role for the male sex hormone testosterone in these mechanisms. Although testosterone was postulated to interfere with the normal development of language areas, Geschwind also suggested that it may lead to the superior development of certain brain areas involved in spatial visualization and visual-motor coordination. If this suggestion is verified, the failure of investigators to obtain evidence for visual-deficit theories of dyslexia would be at least partly explained.

The nonlinguistic theories of dyslexia discussed here have each given rise to remedial techniques designed to correct the cognitive deficits implicated by the theories. For example, visual-deficit theories have spawned remedial approaches designed to improve visual perception, such as optometric training to facilitate binocular coordination, eye tracking and so on. Similarly, cross-modal and serial-memory theories have given birth to a variety of remedial exercises designed to improve these functions, presumably through the direct stimulation of certain brain centers.

If the logical and empirical arguments against the theories supporting these approaches are valid, then the approaches themselves must be seriously questioned. The fact is that research evaluating the efficacy of these approaches for the remediation of reading disabilities has produced no convincing evidence to support them—and a considerable amount of evidence against them. In any case, not enough is yet known about how the brain works to enable anyone to devise activities that would have a direct and positive effect on neurologic functions responsible for such basic processes as visual perception, cross-modal transfer and serial memory.

More conventional approaches to remedial instruction have had greater success, particularly in educational settings equipped to provide dyslexics with the type and amount of help they need. We have found that early remediation of reading difficulties is indicated. It should be based on intensive one-to-one tutoring and a balanced reading program—one that makes generous use of both the holistic/meaning and the analytic/phonetic approaches. The training in reading should be supplemented with enrichment activities to foster language development. Such a program can help a child to develop functional and independent reading skills and so remove him or her from the disabled list. A consensus is emerging among investigators that there is no substitute for direct remedial instruction in reading.



PSEUDOWORDS were shown to two groups of second- and sixth-grade children (both poor and normal readers) who had been given a week's intensive instruction in reading by one of two methods: training in "alphabetic mapping," designed to make them sensitive to the sounds of individual letters, or training in "whole-word meaning," in which pictures of imaginary animals were associated with such nonsense syllables as "zab." For the experiment both groups were shown pseudowords (*top*), made up of novel alphabetic characters, representing the nonsense syllables presented in the whole-word-meaning training. Children learned to say the right nonsense word when it was presented with its pseudoword—just what a child learning to read new words must do. In essence, the alphabetic mappers had to learn to read the pseudowords without benefit of meaning; the children trained in whole-word meaning had to learn to read them without benefit of alphabetic mapping. After several learning trials, subjects were presented with the same characters in reverse order (*bottom*) and asked to learn the names of the new pseudowords. Both the poor and the normal readers who had been trained in alphabetic mapping made very few reversal errors, whereas children trained in whole-word meaning made many such errors.

The Structure of Poliovirus

The virus renowned for its devastating effects has become a model for investigating the molecular links between form and function. Analysis of its structure will enlarge the scope of viral research

by James M. Hogle, Marie Chow and David J. Filman

Forty years after its peak in the Western world, poliovirus leads a double life. It is still a deadly threat in impoverished countries, where, among unvaccinated populations, paralytic polio strikes as many as two million people each year. Yet in wealthier nations the virus has become a valuable scientific tool. Research launched during vaccine development in the 1950's divulged many of its secrets: its host range, its target sites for infection and its life cycle are all well known. As a result poliovirus has become a model system for understanding other viruses.

Until recently, however, it was not possible to study properties of the virus that were directly dependent on its three-dimensional structure. Even though its genes had been thoroughly described, it was difficult to determine, for example, how the simple organism attaches to and enters susceptible cells, how it assembles or what parts of the coat of the virus alert the immune system. Crucial elements in the virus's infectious habits remained out of reach because no one knew what poliovirus looked like in detail.

We now know the intricacies of poliovirus structure. Five years of data collection, millions of calculations and extensive collaboration between our laboratories at the Research Institute of Scripps Clinic and the Massachusetts Institute of Technology have culminated in the construction at Scripps of a three-dimensional atomic "map" of the virus. The maps are based on information provided by X-ray diffraction, a powerful imaging technique we had to customize to accommodate the peculiarities of poliovirus structure. Models derived from these maps offer an unprecedented opportunity to understand the physical bases for many viral functions, and they have already helped to resolve questions raised by earlier investigations.

Perhaps the most practical aspect of our studies is the insight they may lend to research with other, still untamed viruses. A precise understanding of poliovirus structure could further vaccine development and antiviral therapy by elucidating the relation between the structure of a virus and its ability to cause disease. Such information could have direct applications within the poliovirus family, which includes the microbes that cause colds, hepatitis *A* and foot-and-mouth disease. Detailed structural studies have also revealed intriguing similarities between animal viruses and plant viruses that are biologically different.

Our research with poliovirus built on two disparate disciplines, virology and crystallography. Together these disciplines have provided a foundation of knowledge regarding viruses in general. Viruses are very small, very simple organisms that invade cells and usurp their metabolic machinery. A virus directs the unwitting host cell to abandon its purpose and start reproducing viruses instead. The host makes the raw material for the next generation of virus particles and assists with virus assembly, and the new particles go on to infect other cells. In the case of poliovirus infection this activity leads to the metabolic collapse and death of the host cell, but not all viruses kill their hosts.

Viruses vary considerably in their degree of complexity. Some of them consist only of a protein shell or coat that encloses the virus genome: the full complement of hereditary material. Depending on the virus type, the shell can be spherical, bullet-shaped or rod-shaped; its genome can be in the form of one or more segments of DNA or RNA. Some viruses have an envelope surrounding the protein coat that resembles the cell membrane of higher organisms and usually helps the virus to enter and leave host cells.

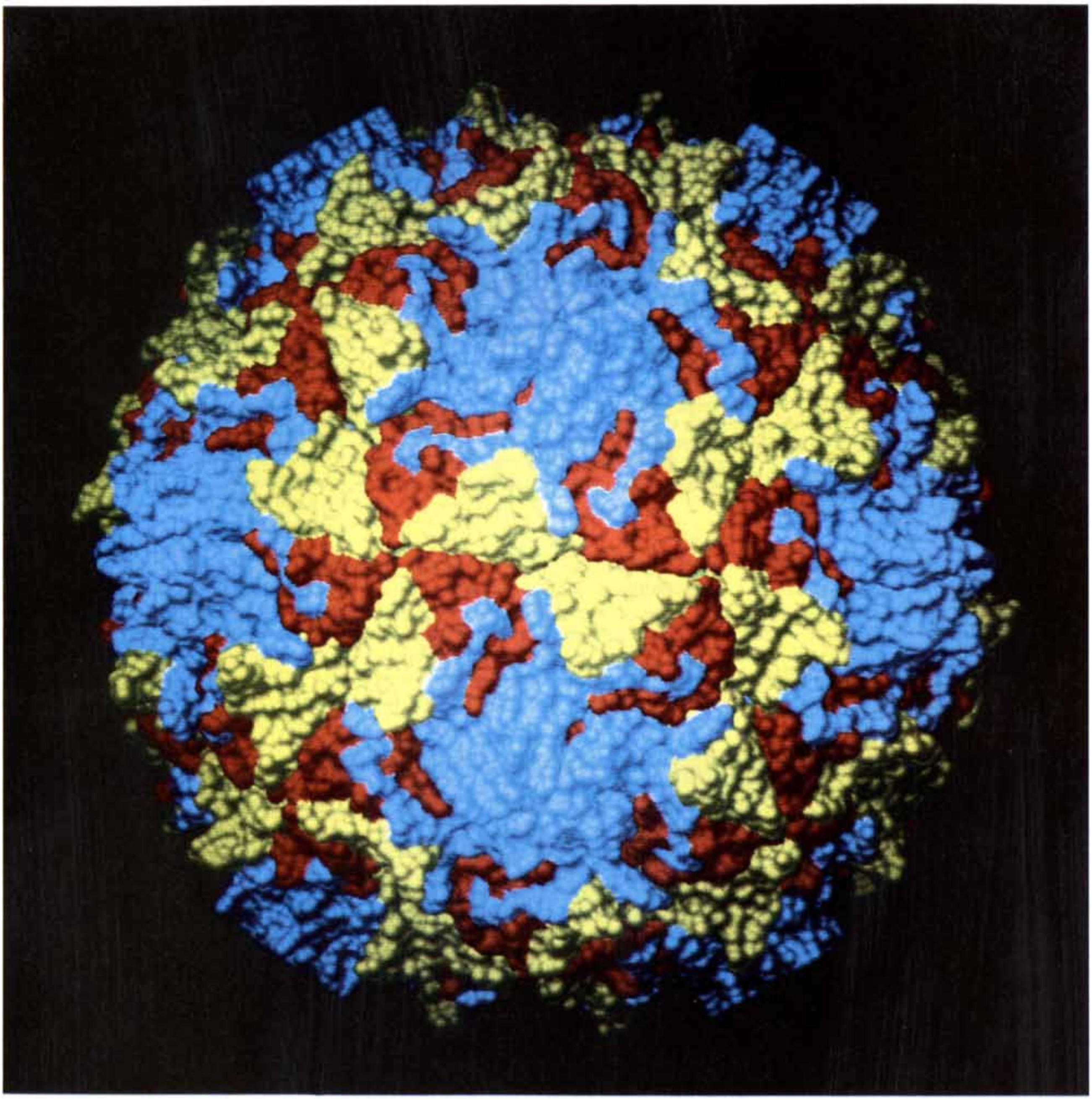
Poliovirus is among the simplest of small spherical viruses, built entirely from protein and RNA. Karl Landsteiner first linked the pathogen with paralytic polio in 1908; by the late 1940's the disease was crippling 20,000 people in the U.S. every year. In 1949 John F. Enders, Thomas H. Weller and Frederick C. Robbins of Children's Hospital in Boston demonstrated that the virus could be grown in cultured cells in the laboratory, thereby liberating poliovirus research from the constraints of primate studies. This critical development enabled two independent investigators, Jonas Salk and Albert Sabin, to introduce vaccines in 1956 and 1961. Subsequent advances in molecular biology enhanced the early characterizations, and poliovirus became the archetype for an entire family called the picornaviruses, or small RNA viruses.

About the same time that the Salk vaccine was introduced in the U.S., James Watson and Francis Crick made a seminal proposal regarding viral structure. They pointed out that the limited coding capacity of the tiny virus genome would be best exploited if a virus coat were composed of many copies of just a few proteins, and that this construction should lead to highly symmetrical configurations. In 1962 D. L. D. Caspar, then at the Children's Cancer Research Foundation in Boston, and Aaron Klug, then at the Medical Research Council Laboratory of Molecular Biology in Cambridge, England, suggested that this symmetry would be icosahedral: a "spherical" virus would in fact have a faceted surface like that of a soccer ball. Caspar and Klug also demonstrated how such icosahedral particles could be assembled from certain specific multiples of 60 subunits.

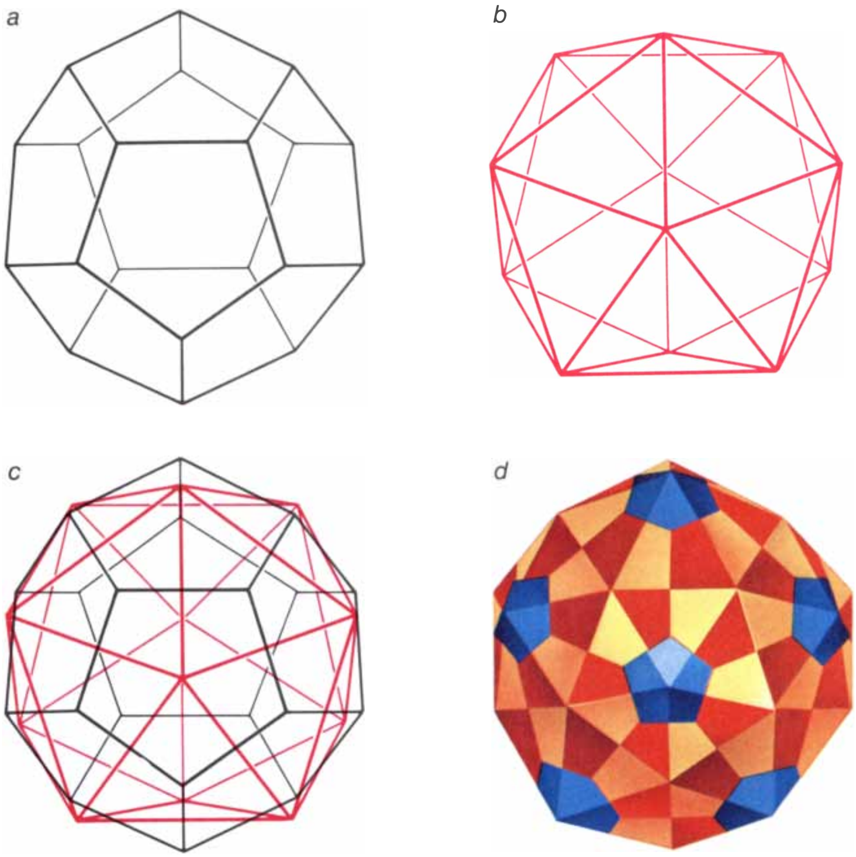
Speculation regarding virus structure could, however, be only as precise as the tools available for its verification. Caspar and Klug based their assumptions on data from electron microscopy and X-ray diffraction. Of the two methods, electron microscopy is the more direct; it follows many of the same principles as light microscopy, except that a beam of electrons rather than a beam of light strikes the sample, and a magnetic lens rather than a glass one focuses the scattered beams to reconstruct the image. Yet even with today's most sophisticated sample-preparation and image-processing techniques, electron micrographs of poliovirus do not attain resolutions powerful enough to distinguish the individual proteins in the virus coat, much less to locate the atoms that make up the proteins.

In any microscopy technique the resolution, or clarity of detail, is limited by the wavelength of the radiation. Short wavelengths can generally discern more detail than long wavelengths. Typical laboratory sources generate X rays with wavelengths of about 1.5 angstrom units (one angstrom is equal to one-tenth of a billionth of a meter). These wavelengths are several orders of magnitude shorter than those of visible light, about the size of the space between atoms in organic molecules. The manner in which a sample causes X rays to diffract, or scatter, is determined by its atomic structure. Hence X rays are ideal for distinguishing individual atoms in organic samples.

Unfortunately two serious difficulties impede visualization with X-ray microscopy in the life sciences. First, biological materials scatter the X rays weakly, making intensities difficult to



COMPUTER-GENERATED MODEL of poliovirus displays the protein subunits that make up the virus shell. The shell contains 60 copies of each of four proteins; three appear on the shell surface and are indicated in blue, red and yellow. The fourth virus-coat protein lines the inside of the shell and is not visible from the exterior. This image, in which poliovirus is enlarged about five million diameters, marks the culmination of five years of research and the arrival of in-depth structural investigations of the virus.



GEOMETRIC MODELS illustrate the underlying symmetry of poliovirus. A dodecahedron (*a*) superposed on an icosahedron (*b*) creates the faceted surface evident in poliovirus structure (*c*). Axes of fivefold symmetry occur where five red lines converge, threefold axes wherever three black lines meet. Projections at the threefold and fivefold axes are also evident in the actual virus structure, which is represented schematically in figure *d*.

detect. Second, lenses for reconstructing the diffracted rays at high resolutions do not exist. The scattered rays can be recorded on photographic film, where they form patterns of dots arranged in concentric circles. Without alternative methods to focus and amplify the rays, however, these diffraction patterns cannot be interpreted.

One way to enhance a weak signal is to induce the sample to crystallize. If a trillion identical virus particles form a stable crystal, the repetitive array of atoms scatters X rays in concert, increasing the intensity of each reflection to detectable levels: hence the "crystal" in "X-ray crystallography."

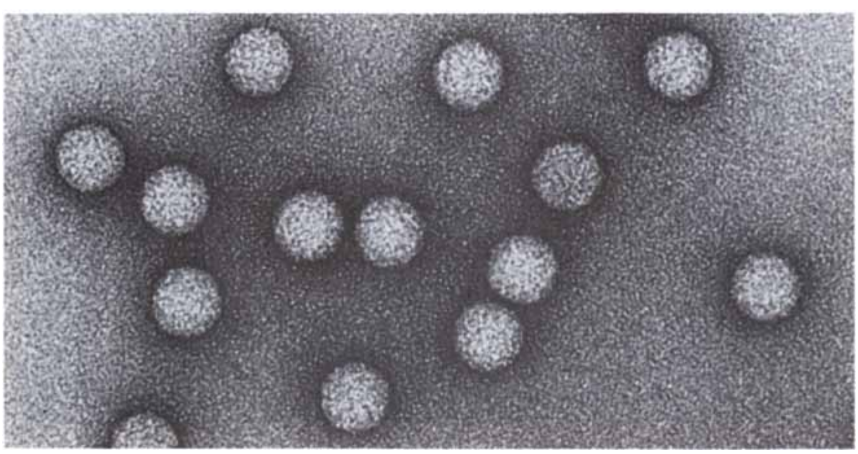
Then a computer can step in as the X-ray-diffraction "lens." If a molecule has fewer than 100 atoms, a computer can reconstruct its image by analyzing statistical relations in the crude diffraction data. (Jerome Karle and Herbert A. Hauptman won the 1985 Nobel prize in chemistry for their role in developing this method.) These statistical techniques are insufficient, however, for proteins, which have thousands of atoms, or for viruses, which have millions of atoms. In these cases "heavy atoms" such as mercury or gold must be introduced into the crystal lattice. These atoms change the intensities of the diffracted X rays in ways that provide additional information for the focusing process.

Provided with the additional data, the computer can carry out its analysis. It then constructs an "electron-density map," a three-dimensional diagram of the distribution of electrons within the crystal and thus a high-resolution guide to the molecules that form the crystal. Interpretation of the electron-density map gives rise to a refined and recognizable three-dimensional atomic model.

Crystallographic analysis of poliovirus dates back to the mid-1950's, when virologists at the University of California at Berkeley found they had accidentally allowed some poliovirus samples to crystallize. The Berkeley group sent the crystals to Birkbeck College London for X-ray analysis. There Rosalind Franklin, and after her death Klug and his colleague John Finch, collected the data that would lead to Caspar and Klug's hypothesis that spherical viruses had icosahedral symmetry.

Crystallography research in the late 1950's was overwhelmed by several technical problems, among them the sheer magnitude of numbers. Whereas structural determinations of small molecules require only a few hundred measurements, the structures of medium-sized proteins (of which some 300 have been solved to date) typically take tens of thousands; structures as large as viruses entail millions of measurements. During the 1960's and 1970's advancements in crystallography were limited by the power and cost of computing.

The symmetry of poliovirus can itself serve as a powerful independent source of information because it imposes considerable constraints on processing data to form an image. David M. Blow and Michael G. Rossmann had examined ways to exploit



ELECTRON MICROGRAPHS of poliovirus do not achieve the resolution needed for precise structural analyses. Yet these were the best images available before X-ray crystallography disclosed virus structure. The particles here were enlarged 300,000 diameters.

this symmetry in 1961, and their efforts proved somewhat successful in studies using low resolutions. But attempts to apply their methods at high resolutions faltered when the calculations themselves became unmanageably large.

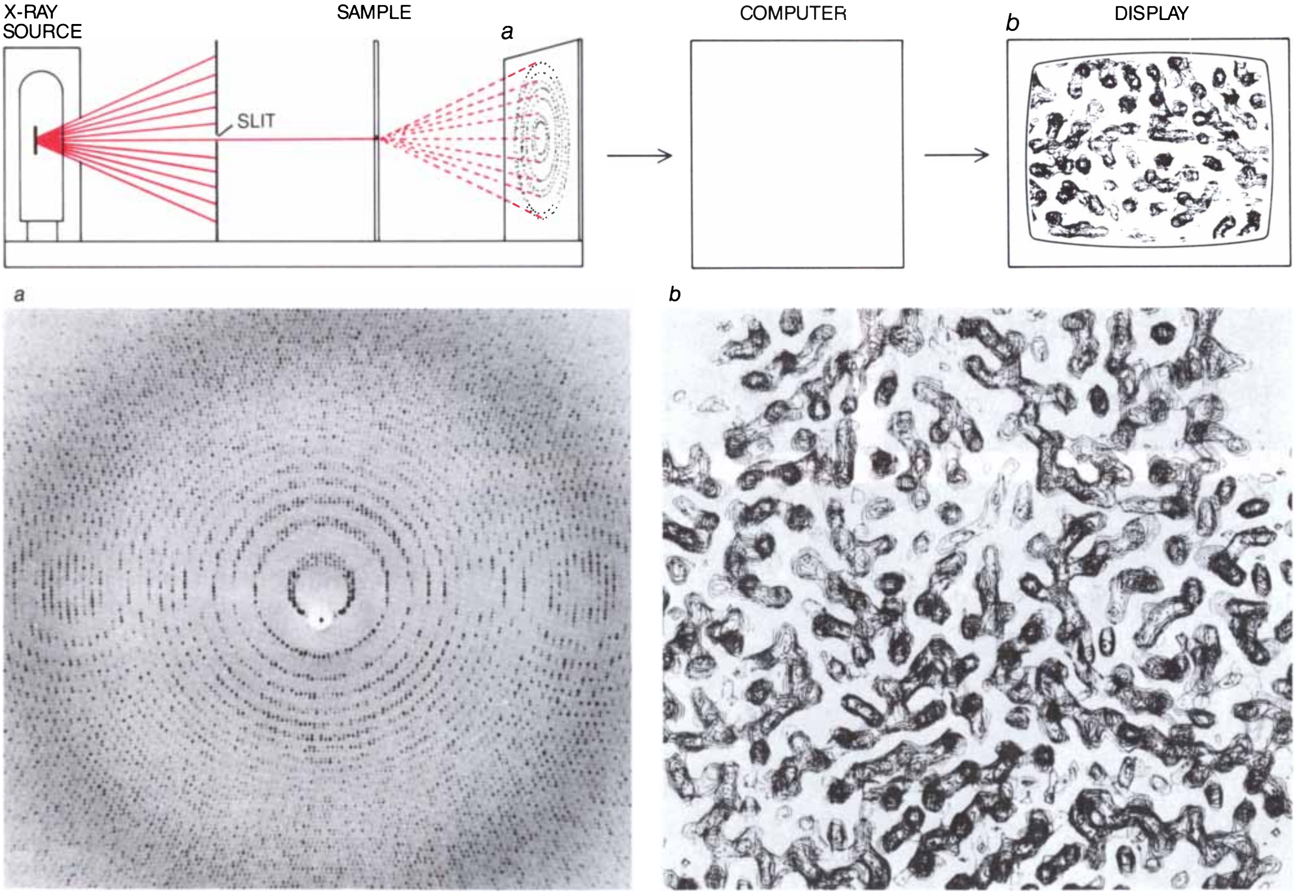
In 1974 Gerard Bricogne, then at the Medical Research Council laboratory in Cambridge, developed an elegant technique for including symmetry information in the imaging process. On the basis of this research he developed a simplified algorithm that enabled computers to solve very large symmetrical structures at high resolution. His procedure was instrumental in determining the first high-resolution structure of an intact virus, tomato bushy stunt virus (TBSV). The determination was reported in 1978 by Stephen C. Harrison and his colleagues at Harvard University.

Several other plant-virus structures have been solved since then, including southern bean mosaic virus (SBMV), satellite tobacco necrosis virus (STNV) and turnip crinkle virus (TCV). Although this research has advanced the understanding of the architecture and assembly of simple viruses, it has been less successful at linking function with structure because so little is known about the biology of plant viruses.

In contrast, the wealth of information about the biology of the animal viruses, and of polio and other picornaviruses in particular, makes them very attractive candidates for structural studies. Soon after the first plant-virus structures were reported, we therefore began to study a common laboratory strain of poliovirus called the Mahoney strain of type 1 poliovirus. From this strain we obtained crystals that turned out to be identical in appearance and crystallographic properties with those grown accidentally at Berkeley and characterized by Finch and Klug at Birkbeck more than 25 years earlier.

The acquisition of data from these crystals and the subsequent structure determination represented a formidable undertaking. We collected and processed almost five million measurements from a total of 84 crystals in the course of our investigations. Even though the poliovirus is no larger than most plant viruses, our crystals produced more numerous and overlapping X-ray reflections, creating problems that required the development of new techniques and novel computer software.

In spite of these considerable difficulties, we eventually managed to reach a resolution of 2.9 angstroms—that is, our maps could discriminate between objects 2.9 angstroms apart. Our resulting electron-density map was remarkably clear. We had no difficulty identifying each of four protein chains in the contours of the map. Owing to the work of David Baltimore and his colleagues at M.I.T., we already knew the linear sequences of the linked chemical groups, called amino acids, that constitute these protein chains. We were able to match the mapped chains with the molecular sequences by identifying characteristic amino acid side groups in our map. We now had a way to correlate our three-dimensional structure with information provided by years of biological



X-RAY CRYSTALLOGRAPHY involves several steps, rendered schematically in the drawing. X rays pass through a slit that eliminates stray beams; the focused beam strikes a crystalline sample, which scatters the rays in a pattern defined by the sample's structure. Photographic film records the pattern, an example of which is shown at the lower left (*a*). By analyzing sets of these diffraction patterns, a computer can calculate the distribution of electrons, and therefore of atoms, in a sample. Part of a computer-generated electron-density map is shown at the lower right (*b*). Such maps serve as guides for models such as the one on page 43.

studies. Our model gave physical substance to the chemical evidence supplied by previous picornavirus investigations. Altogether determining the structure took almost five years.

Other researchers had found that the poliovirus coat is composed of many copies of just four proteins, called VP1, VP2, VP3 and VP4. We showed that of the four, VP1, VP2 and VP3 are large and rather similar in structure. Each of these three virus proteins contains an eight-strand core called a beta barrel and can be distinguished by the characteristic loops and tails that decorate the barrel. The cores pack together to form the continuous protein shell, providing a structural scaffolding for the virus; the tails coat the inside of the virus shell and the loops decorate the virus surface. The loops appear to play a dominant role in interactions with elements outside the virus. VP4, only a fourth the size of the other three molecules, is a kind of detached tail of VP2; it also lines the inside of the protein shell.

The virus coat consists entirely of 60 copies of each of these proteins. The coat assembles from 12 compact aggregates, called pentamers, which contain five of each of the coat proteins. Each pentamer is shaped like a molecular mountain: VP1 clusters at the peak, VP2 and VP3 alternate around the foot and VP4 provides the foundation. When the pentamers come together to form the virus coat, the VP1 mountain peaks at the fivefold axis of symmetry and another smaller and flatter protrusion, like a mesa, occurs at the threefold axis, where VP2 and VP3 meet. Although the RNA genome is not visible in our experiments, calculations indicate that it must be packed into the central cavity very densely.

Thus the poliovirus map has revealed the exact folding patterns for each of the four coat proteins. Because folding defines the shape of each molecule and the relative placement of important chemical groups, it is inextricably tied to the biological function of the protein. For instance, by looking closely at specific parts of the proteins that make up the virus, we were able to suggest how the manufacture of virus proteins in the cell is coupled with virus assembly.

When poliovirus infects a cell, it quickly monopolizes the biosynthetic capabilities of its host: two or three hours after infection, its RNA is directing the synthesis of virus proteins. All four of the coat proteins are initially contained in one long stretch of amino acid chain called P1. P1 is cleaved by enzymes into three proteins: VP0, VP1 and VP3, in a step that seems to be intimately associated with the assembly of these proteins into pentameric aggregates.

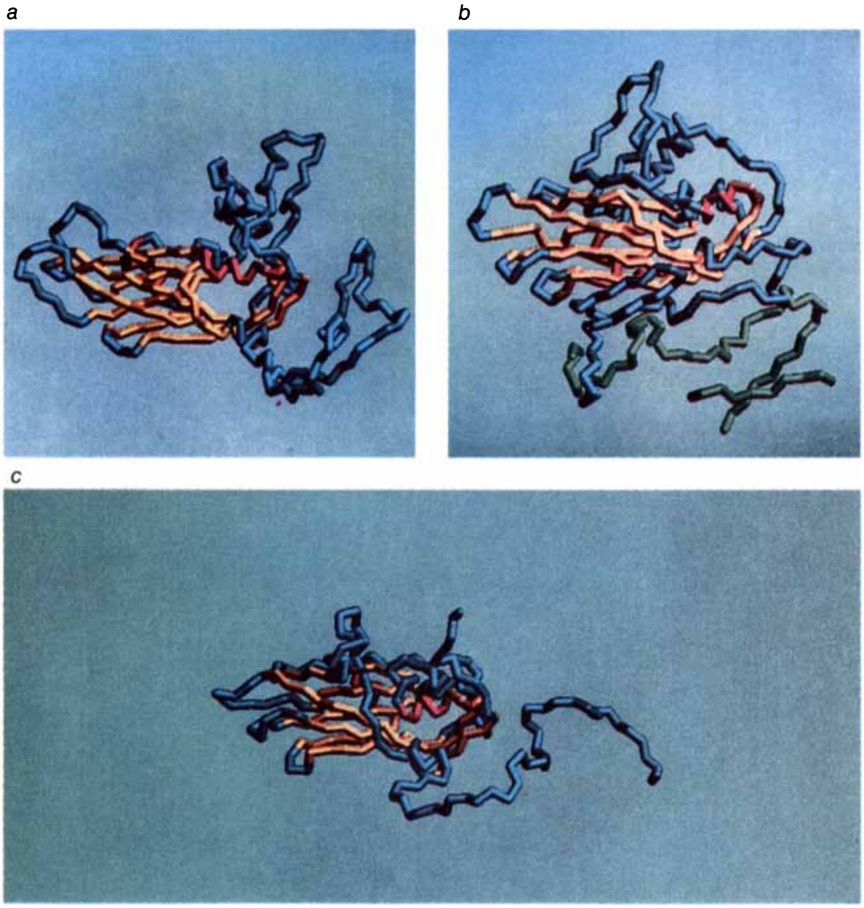
The virus structure suggests how cleavage and the assembly of pentamers might be related. Our model shows that five of the VP3 ends freed by these cleavages become intertwined in a twisted helical network. In forming this stable and compact network during virus assembly, these VP3 tails may encourage protein subunits to associate in the correct pentameric configuration. Thus molecular rearrangements made possible by cleavage serve to direct pentamer formation.

Next the pentamers organize into a protein coat with 60 copies each of VP0, VP1 and VP3. It is not known whether the pentamers encapsulate RNA when they form the shell or whether they construct an empty capsid that subsequently incorporates the RNA. These last steps can easily be reversed, so that the growing shell is an unstable structure. In the final assembly step VP0 splits into VP2 and VP4 to yield a mature virus. Once this irreversible "maturation cleavage" has taken place the virus can no longer be disassembled except by irreparable disruption.

Our structural findings have also contributed clues that further knowledge of this final assembly step. The model revealed that the VP2–VP4 cleavage site is isolated inside the virus coat, where it is inaccessible to external enzymes. Therefore the cleavage is probably catalyzed by the virus protein itself. The spatial arrangement of chemical groups on the inside of the shell, made apparent by our studies, is such that they could effect such a cleavage in the presence of RNA.

Our work also revealed some unexpected insights into viral evolution. We found some predictable similarities between poliovirus and rhinovirus 14, another picornavirus whose structure was reported by Rossmann and his colleagues at Purdue University in collaboration with Roland R. Rueckert of the University of Wisconsin at Madison. Rossmann's group published its report shortly before we presented our findings. The similar-

a
b
c



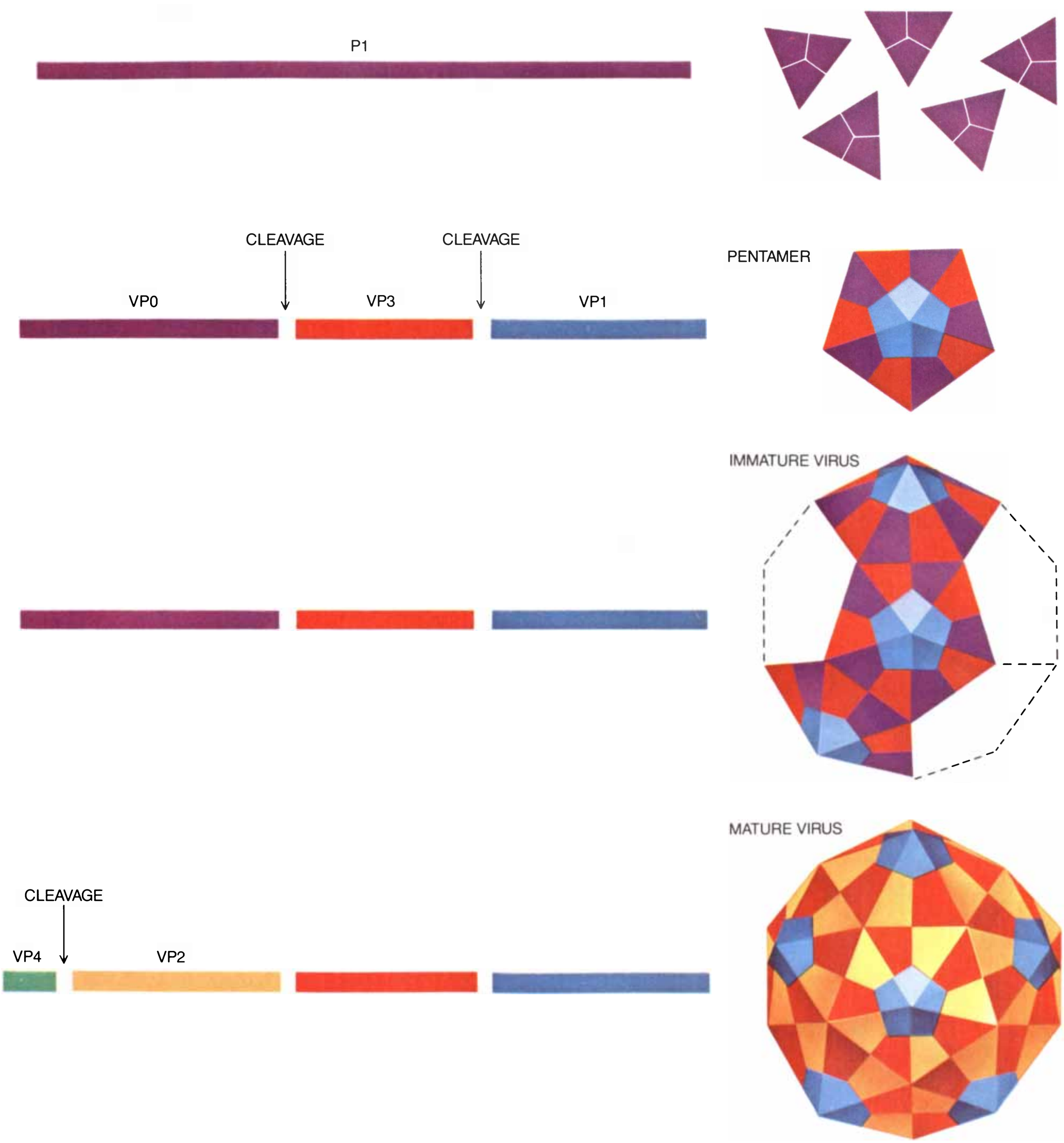
COAT PROTEINS are the building blocks of poliovirus. Each protein consists of one amino acid chain folded into a characteristic shape. Three of the four proteins, VP1 (*a*), VP2 (*b*) and VP3 (*c*), are similar in size and have closely related cores (*red and yellow*). Other parts of the proteins differ (*blue*). The fourth protein, VP4, is designated by the green chain in *b*. VP4 is the smallest virus protein and acts like an extension of VP2.

ities the viruses share, which occur both in the protein subunits and in the packing pattern of the shell, correspond to similarities in their genomes.

Less predictably, the coat proteins of the picornaviruses also resemble those of the icosahedral plant viruses: the coat proteins of the plant viruses have identical folding patterns and similar overall shapes. Moreover, the packing arrangement of the polio and rhinovirus coat proteins is strongly reminiscent of (although not identical with) the way that 180 subunits pack in the shells of the plant viruses TBSV, SBMV and TCV.

These structural similarities point to an evolutionary relation between the picornaviruses and the plant viruses. The concept that plant viruses and animal viruses may be related is not surprising. What is surprising is that these similarities have persisted even though the viruses belong to vastly different families and differ greatly in the details of their molecular biology. The finding raises the intriguing possibility that viruses may have evolved by stealing bits of genetic material from host cells or other viruses during brief encounters with the genomes of the other organisms.

In addition to providing informa-

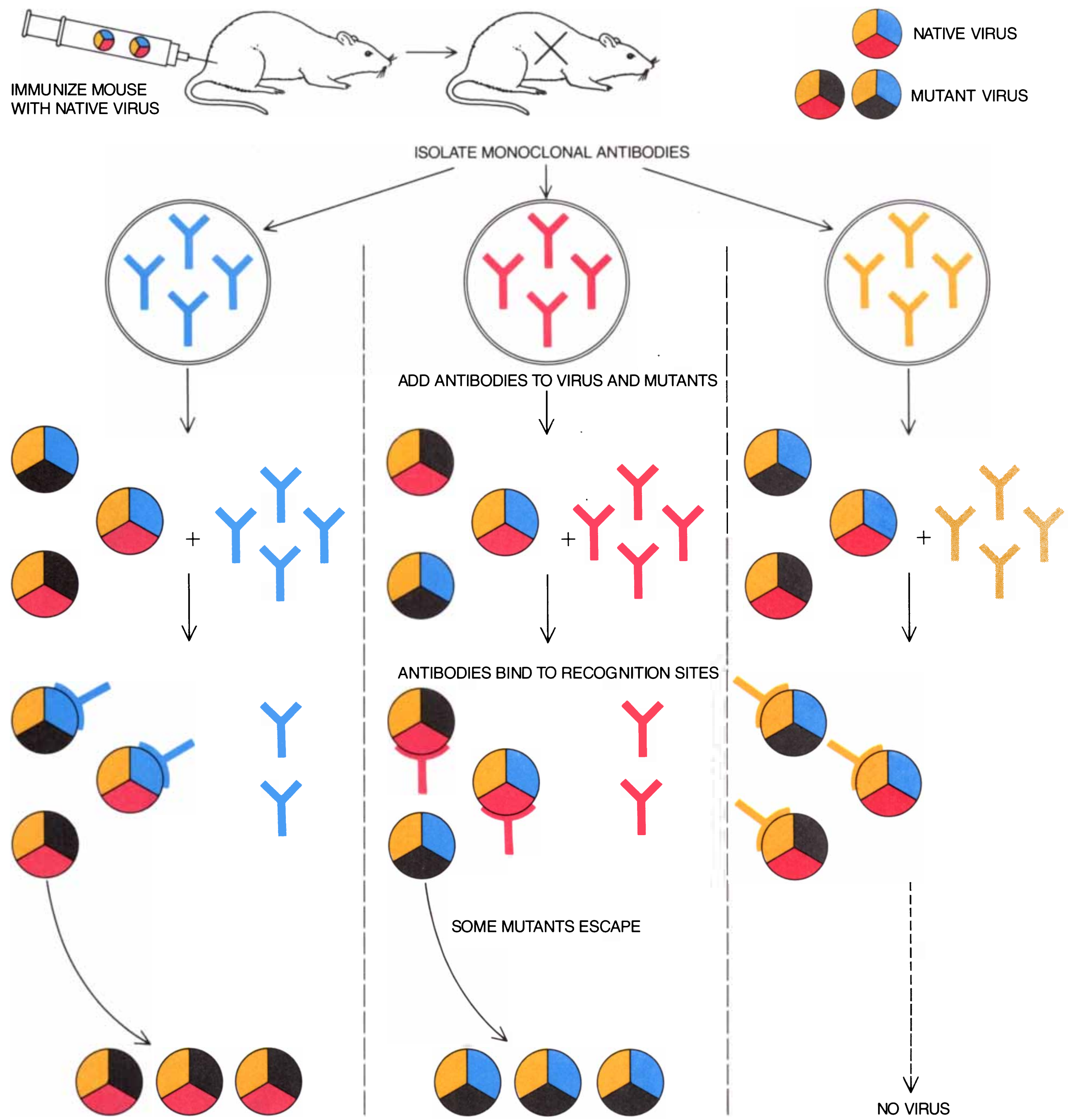


STEPS IN VIRUS ASSEMBLY parallel events in the cleavage of virus proteins. All four coat proteins originate from a single chain called P1 that folds into a three-dimensional subunit. When enzymes cut the chain to yield VP1, VP3 and VP0, rearrangements of the protein "tails" encourage five of the subunits to group together in a pentamer configuration. Then 12 pentamers aggregate to form the coat. The coat structure remains unstable, however, until VP0 is cleaved into VP2 and VP4 in the final assembly step.

tion about the physical properties and assembly of poliovirus, the structural model is giving clues to the process by which the immune system recognizes and neutralizes the virus. Now that a three-dimensional view is available, it is possible for the first time to visualize directly the structure of an antibody-recognition site on an intact virus. These sites are the "targets" for attack by antibodies, the immune system's army. The spatial arrangement of the sites might yield fundamental insight as to how they function.

Antibody-recognition sites are often identified by isolating monoclonal escape mutants: genetically altered forms of poliovirus that escape neutralization by monoclonal antibodies. Usually these mutants survive because they contain a mutation that changes a recognition site, thereby preventing antibodies from attaching to the coat of the virus. By determining the position of the mutation in the viral RNA, one can identify the altered amino acid and track down its position on the poliovirus coat.

Many laboratories have been isolating these monoclonal escape mutants,

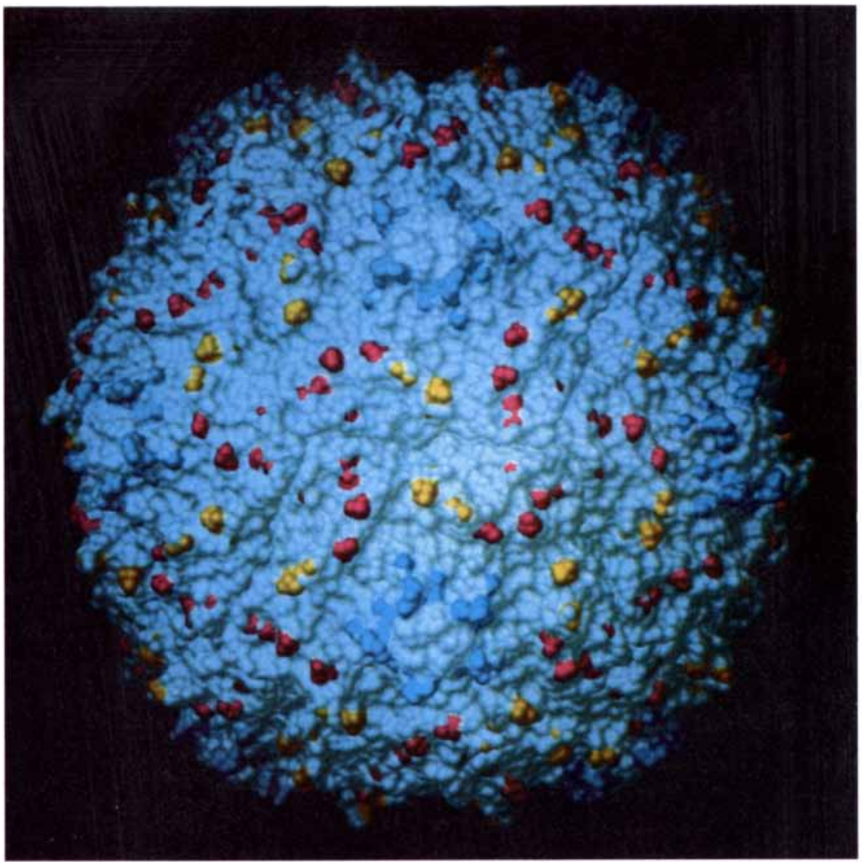


ISOLATION OF ESCAPE MUTANTS is a procedure critical to studies of the immune response to poliovirus. Mice are injected with native, or normal, poliovirus. Cells that make antibodies thereupon proliferate in the mice. Each antibody is designed to interact with a specific part of the virus coat; this specificity is represented in the drawing by different colors. The antibody-producing cells are collected, fused with tumor cells and segregated to create immortalized cultures that synthesize monoclonal antibodies. When native poliovirus and poliovirus mutants are subsequently exposed to these antibodies, most become inactivated by antibody attachment. But some mutants may exhibit changes in coat proteins that prevent the antibodies from attaching; these mutants escape. It is possible to identify the sites that antibodies recognize by comparing escape mutants with native poliovirus.

including Geoffrey C. Schild and his colleagues at the National Institutes for Biological Standards and Control in London, Eckard Wimmer and his colleagues at the State University of New York at Stony Brook and Radu Crainic and his colleagues at the Pasteur Institute. We have also selected and characterized several variants in an ongoing collaboration with Anne Moser of the University of Wisconsin.

We used the poliovirus model to see for the first time how clusters of these "point" mutations make up parts of the same antibody-recognition site on an intact virus. We knew that many recognition sites on antibodies are quite extensive: they measure some 30 angstroms in diameter, which is a relatively large area on the surface of the virus. Because the surface of a single amino acid side chain is only a few angstroms in diameter, it is clear that a site must be composed of several amino acids. Hence we predicted that mutations within 30 angstroms of one another might be parts of the same site.

We identified three such clusters of mutations, each consisting of protein segments that are distant from one another in the linear sequence of the protein but close together in the three-dimensional structure. Two of the clusters include portions of more than one of the coat proteins. Recent immunological experiments have confirmed that these clusters of mutations do in fact form parts of three antigenic sites. Specifically, two or more mutants that lie in the same cluster each allow the virus to escape neutralization by the same monoclonal antibody. Thus predictions based solely on spatial distribution in the virus model have now been empirically verified.



THREE RECOGNITION SITES on poliovirus were identified, each a composite of many individual amino acids. The sites, which act as targets for the immune system, are shown here in red, dark blue and yellow. They recur throughout the virus coat owing to its repetitive structure. Without a three-dimensional poliovirus model, the spatial organization of many amino acids within a single recognition site would not have been apparent.

The mutation sites identified by monoclonal escape variants are all located in conspicuous loops on the surface of the virus, areas that are readily accessible to antibody binding. This finding suggests that animal viruses may decorate their outer surfaces with loops that can easily accommodate mutations in order to escape the immune system's surveillance. Rhinovirus 14 exhibits a similar site distribution, as do antibodies raised against "synthetic peptides" (man-made bits of protein that correspond to parts of the virus-coat proteins).

These peptides, however, have also identified antigenic sites within the virus's interior. It is hard to explain how these buried sites could be accessible to antibody binding unless one postulates that, at some point in its life cycle, the virus undergoes large conformational changes that expose the sites.

Although we are encouraged by the information we have already gleaned from our structural studies, their impact should reach beyond these preliminary interpretations. For instance, the conformational dynamics that appear to play a part in the immunogenicity of poliovirus might well be important to host-cell recognition, cell attachment and uncoating of the viral RNA inside the cell. Such involvement has been implied by the fact that all 60 copies of VP4, a protein confined to the interior of poliovirus, are lost when the virus attaches to and penetrates susceptible cells.

The poliovirus model might also become a key to the relation between a virus's structure and its ability to cause disease. Although certain strains of poliovirus grow readily in the central nervous system, others do not. Vincent R. Racaniello of Columbia University has shown that in at least one strain the virus's ability to cause paralytic disease can be intimately linked to the nature of the coat proteins. If investigators can use the poliovirus model to define the structural features that control entry of poliovirus into the central nervous system, they might be able to apply that knowledge to design a new generation of polio vaccines and to defeat other pathogens that infect the central nervous system.

In fact, we hope that a better understanding of the structural aspects of the poliovirus life cycle will suggest many ways to combat related viruses against which there is now no defense. By systematically altering the structure of the virus with genetic-engineering techniques, investigators can gain a working familiarity with the correspondence between structural features and functional properties. Antiviral drugs could be designed more effectively if the role of structure in viral functioning were better understood. And poliovirus could become a vehicle for other vaccines if sites were found where bits of related viruses (for example hepatitis *A*) could be grafted onto the protein coat. Just a few decades after it brought devastation to the Western world, poliovirus may yet lead the battle against unconquered viral infections.

Cooling and Trapping Atoms

Atoms are slowed and cooled by radiation pressure from laser light and then trapped in a bottle whose "walls" are magnetic fields. Cooled atoms are ideal for exploring basic questions of physics

by William D. Phillips and Harold J. Metcalf

A highly fruitful area of scientific research has traditionally been the study of the intrinsic properties of isolated atoms. In the early part of this century such inquiry led to the formulation of quantum mechanics, one of the cornerstones of modern physics. More recently, precise measurements with atoms have also shed light on other fundamental physical theories, including relativity.

The triumphs born from probing atomic systems have depended largely on the ability to make precise measurements. A deeper understanding of the structure of matter at the atomic level requires that measurements be made with even greater precision. Unfortunately the necessary precision cannot be easily achieved: in solids and liquids one cannot isolate individual atoms from the effects of their neighbors, and in gases the random thermal motion of atoms makes highly precise measurements difficult.

The continued need for ever more precise measurements has led to the development of many techniques for overcoming the effects of thermal motion. Simply reducing the effects of thermal motion is not sufficient for the most demanding measurements; ultimately it becomes necessary to reduce the thermal motion itself. Methods developed over the past few years provide the required means. Atoms can now be cooled by shining laser light directly on them. The radiation pressure exerted by the laser light can be exploited to push on the atoms in order to slow them down. Once the atoms have been cooled, they can be trapped, or confined to a limited region of space. In our laboratory we trap atoms in a bottle whose "walls" are electromagnetic fields rather than material substances.

Cooled or trapped atoms are ideal for exploring fundamental questions of physics. Early in 1986, for instance, cooling and trapping were employed to observe for the first time a fundamental quantum-mechanical process: the transition of a single atom from one discrete energy level to another (a quantum jump). In this case the trapping was relatively easy because the atom had been ionized, or stripped of one of its electrons. Recent techniques have made possible the trapping of even electrically neutral atoms.

In the future, laser-cooled atoms will certainly facilitate spectroscopic measurements, perhaps leading to substantial improvements in such areas as atomic clocks and measurements of fundamental constants. Cooling and trapping will enable investigators to look at collisions between atoms in detail and better understand the way chemical bonds are formed. Laser cooling and electromagnetic trapping could also be employed to manipulate atoms made of antimatter. Since antimatter annihilates with ordinary matter on contact, ordinary material walls cannot contain it.

At high enough densities certain atoms might even undergo a fundamental type of transition called a Bose condensation. Such a condensation is predicted by quantum mechanics, which holds that there are two types of fundamental particles called fermions and bosons. Fermions, which include electrons, protons and neutrons, cannot be in the same quantum state; in contrast, bosons, which include some atoms (such as the hydrogen atom), can occupy the same quantum state. In a sufficiently cold and dense collection of bosons a major fraction of the particles will be in the same, lowest-energy quantum state. The eagerly sought observation of Bose condensation could well occur in traps populated by laser-cooled atoms.

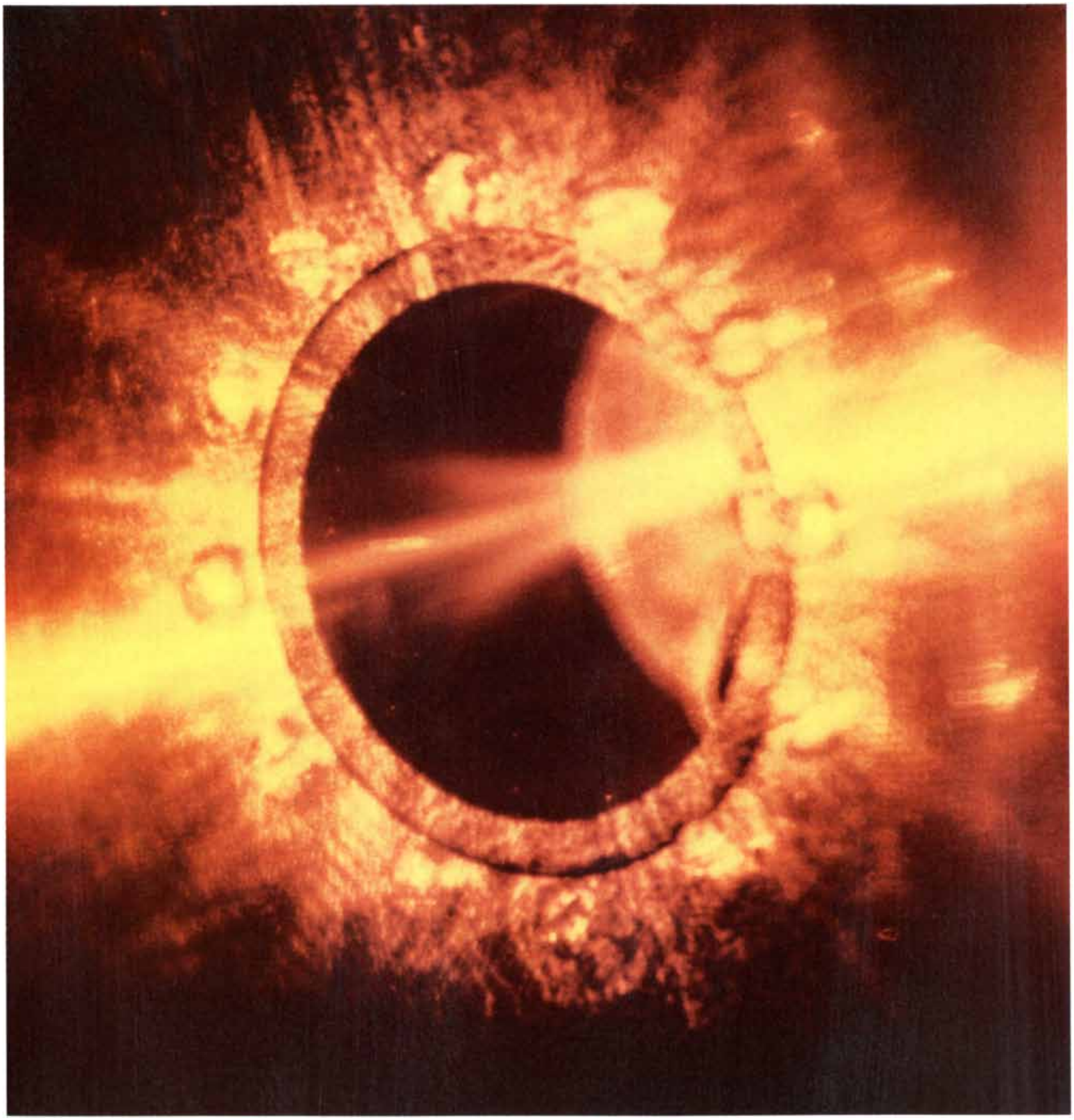
Why are atoms always in motion? The kinetic theory developed during the past century shows that gases are collections of atoms or molecules whose average kinetic energy (which varies directly as the square of the velocity) is proportional to the absolute temperature. Such classical motion can never be eliminated except at the unattainable temperature of absolute zero, or approximately −273 degrees Celsius. At room temperature, for instance, air molecules move at an average speed of about 500 meters per second, or 1,100 miles per hour. Furthermore, the atoms move at various speeds, mostly between zero and twice the average speed.

Why does thermal motion affect precise measurements of atoms? Each atom in the gas acts like a sharply tuned transmitter and receiver of electromagnetic radiation; an atom can efficiently emit and absorb only certain frequencies of radiation. Since different kinds of atoms emit and absorb different frequencies, each set of frequencies, called a spectrum, serves as a "signature" for a particular type of atom. In other words, identical atoms have the same spectrum and share the same signature. The measurement of spectra, called spectroscopy, allows one to draw conclusions about the fundamental structure of atoms.

Unfortunately the observed spectrum of identical atoms in a gas is smeared out at ordinary room temperatures because of the thermal motion of the atoms. When an atom moves with respect to an observer, its characteristic frequencies appear to shift from the intrinsic ones seen when the atom is stationary. The phenomenon is called the Doppler shift, after the 19th-century Austrian physicist Christian Doppler, who explained a similar effect for sound. The basic phenomenon is familiar to anyone who has heard the sudden change in pitch of the horn of a passing train. It is also the Doppler shift that enables a police radar to determine the speed of a car.

The Doppler shift also occurs when an atom moves with respect to a light source. The atom encounters the crests and troughs of the radiation waves at a higher rate, so that the frequency appears to the atom to be shifted up, or higher. Conversely, if the atom moves away from the source, the apparent frequency is shifted down, or lower. Even in the absence of this kind of Doppler shift, as when the atom moves perpendicularly to the line connecting it to the source, there is always another apparent frequency shift owing to special relativity. (The relativistic effect arises from the fact that an observer sees a moving clock run slower than his own identical clock. One consequence of this is the "twin paradox": a twin who travels in a spaceship is found on his return to have aged less than his earthbound sibling.)

Because of the effects of the Doppler shift, a gas of randomly moving atoms—each identical with the others and each having an identical spectrum—appears to be a collection of atoms all of which have slightly shifted spectra. The spreading in frequencies, while it is only a millionth of the optical frequency, has profound consequences. For example, for many years the Doppler broadening of the optical spectrum of hydrogen hid a small but important frequency shift called the Lamb shift. The eventual discovery of the Lamb shift (by nonoptical meth-



LASER COOLING brings atoms moving at speeds exceeding 1,000 meters per second, or 2,200 miles per hour, to a virtual standstill. Here a laser beam shines into a collimated beam of sodium atoms. The sodium atoms enter from the right of the photograph, the laser beam from the left. The atoms slow down from the radiation pressure of the laser light and come to a virtual halt near the center of the circular opening (which is at the end of a solenoid: a long coil). The slowed atoms spread out into a "skirt."

ods) ultimately led to and confirmed the theory of quantum electrodynamics, which is now thought to be a complete description of the interaction of radiation and matter and the paradigm for all modern field theories.

Yet another motional effect that plagues spectroscopy is transit-time broadening. Since atoms are moving, they do not stay in a region where they can be observed for long periods of time. The limited time available for measurement broadens the spectrum. The higher the atomic velocity, the smaller the observation time and the greater the broadening.

Many methods have been developed to overcome the difficulties imposed on spectroscopy by atomic motion. Spectroscopy that is virtually free of Doppler effects can be achieved by a variety of schemes. Still, there are residual effects that arise from imperfections in the apparatus; furthermore, these schemes do not address the problems of transit-time broadening and relativistic shifts, which limit spectroscopy requiring an accuracy of one part in 10^{11} or better.

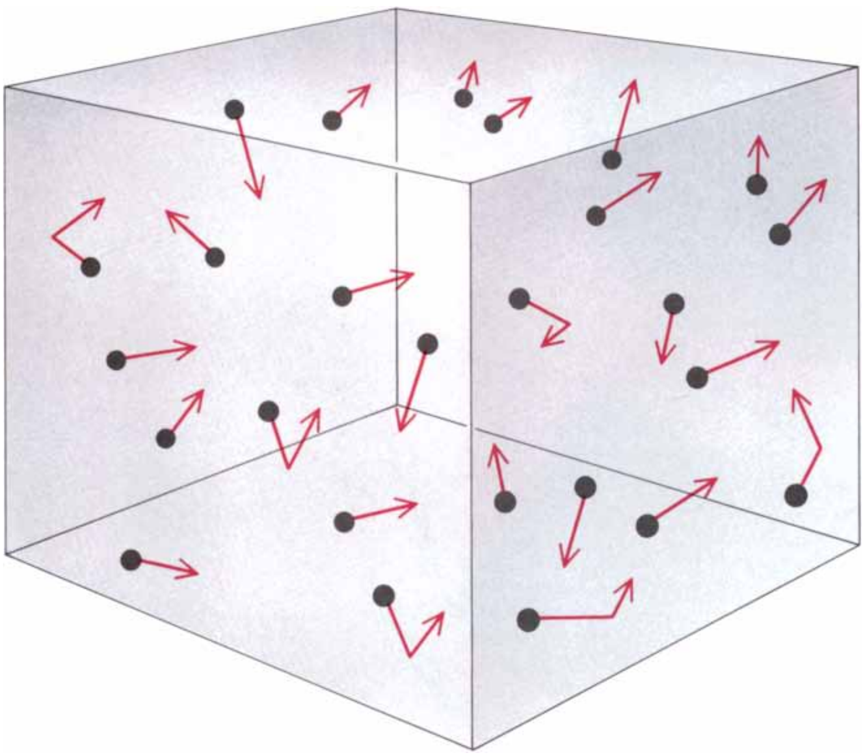
Cooling is the most obvious way to reduce the motion of atoms and minimize such effects, thereby making measurements more precise. One way to accomplish cooling is to have the atoms collide with either the walls of a cold container or another gas of cold atoms. Such cooling works up to a certain point. At low enough temperatures nearly any atom will condense on the container walls or form molecules or clusters with its collision partners. When either happens, the atom is no longer isolated and complicated interactions with its neighbors prohibit accurate measurements.

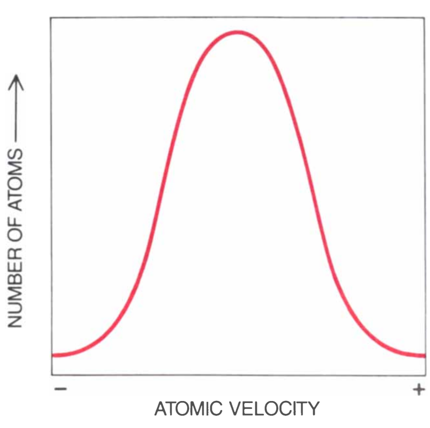
A different kind of cooling—laser cooling—is needed to achieve the low velocities associated with low temperatures without inducing unwanted condensation of atoms. In a sense the highly organized, monochromatic nature of laser light makes the laser a very low temperature system through which atoms can give up heat by their interactions with it. The idea of laser cooling was proposed independently in 1975 by Theodor W. Hänsch and Arthur L. Schawlow of Stanford University for a gas of atoms and by David Wineland and Hans G. Dehmelt of the University of Washington for trapped ions.

Laser light can be exploited to affect atomic motion and cool atoms because it has momentum, and so it pushes against objects that absorb or reflect it. Momentum is a characteristic of motion that can be transferred but not created or destroyed. Matter and light interact and transfer momentum by exchanging discrete packets of light called photons. The number of photons necessary to cool atoms is enormous. In order to bring to rest a single sodium atom traveling 1,000 meters per second, for instance, some 30,000 photons must strike it head on.

To see how laser cooling works, suppose a laser beam shines into a gas of identical atoms at room temperature. Further, suppose the laser beam is tuned to a particular frequency that is lower than one of the intrinsic frequencies at which the atoms emit and absorb radiation. Some of the atoms in their rapid and random motion will therefore have just the right velocity and corresponding Doppler shift to absorb the light strongly, but for the majority of atoms the light will have little effect. In particular, those atoms moving toward the light source "see" it Doppler-shifted closer to their intrinsic frequency, and therefore they absorb light more rapidly. Those atoms will slow down, because the momentum of the light opposes their motion. On the other hand, atoms moving away from the source are less likely to absorb the light, because they "see" the frequency of the laser Doppler-shifted to a still lower value, further away from the frequency they need for absorption. The atoms moving away from the source, which are accelerated when they absorb light, do so much more slowly than the atoms moving toward the source. The net effect is that on the whole the atoms in the gas slow down.

In 1978 workers at the National Bureau of Standards in Boulder and at the University of Heidelberg independently demonstrated laser cooling of trapped positive ions (atoms that have





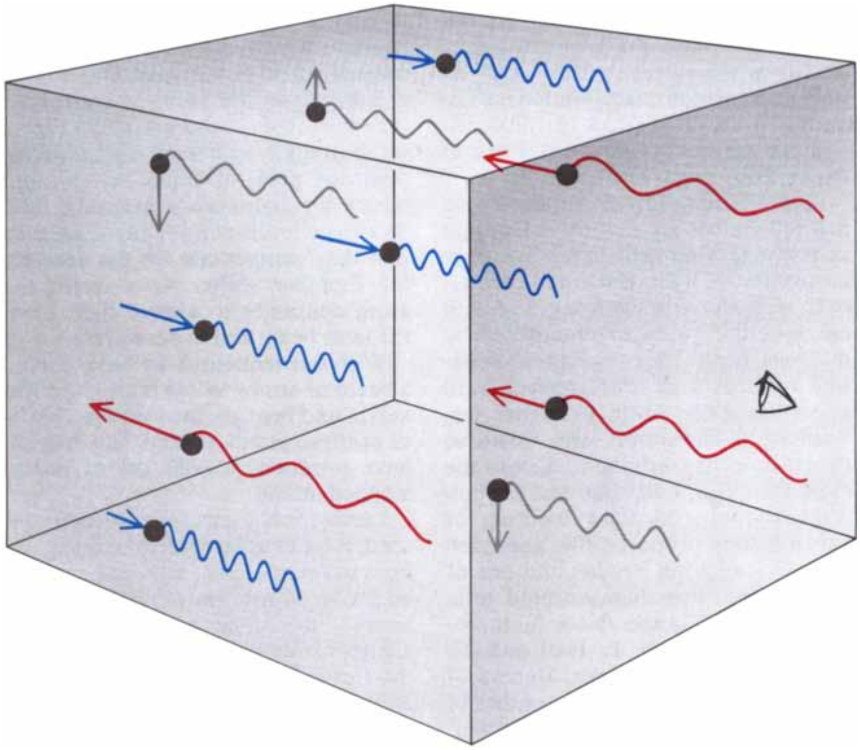
ATOMS in a gas held at room temperature move rapidly, at varying speeds and in different directions (*top*). The atoms collide frequently with one another and, if they are held in a container, with the walls of that container. The distribution of velocities of individual atoms in a gas is given by a bell-shaped curve called a Maxwell-Boltzmann distribution (*left*). The distribution actually shows the number of atoms as a function of the projection of their velocity along a given axis. The higher the temperature of the gas is, the wider the distribution is and the harder it becomes to make precise measurements. One way to avoid such problems is to cool the atoms.

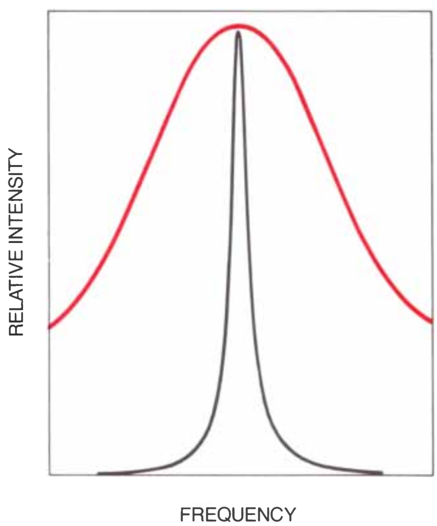
lost one or more electrons and so have a net electric charge). Continued experiments with trapped ions by these and other groups have produced ions cooled to only a few thousandths of a degree above absolute zero, highly accurate atomic clocks and visual observations of single atoms.

We have been working at the National Bureau of Standards in Gaithersburg, Md., to extend the technique of laser cooling to neutral atoms. A key difficulty in cooling neutral atoms is that, unlike electrically charged ions, they are not easily influenced by electric and magnetic fields. Although ions held at room temperature and above are readily trapped in such fields, neutral atoms must first be cooled to temperatures on the order of one degree Kelvin (−272 degrees C.) or below before they can be trapped. Over the past several years our group and other groups at the Institute of Spectroscopy in Moscow, the National Bureau of Standards in Boulder, the AT&T Bell Laboratories, the University of Colorado at Boulder, the University of Bonn, the École Normale Supérieure in Paris, the State University of New York at Stony Brook and the Massachusetts Institute of Technology have successfully developed approaches to cooling neutral atoms with laser beams.

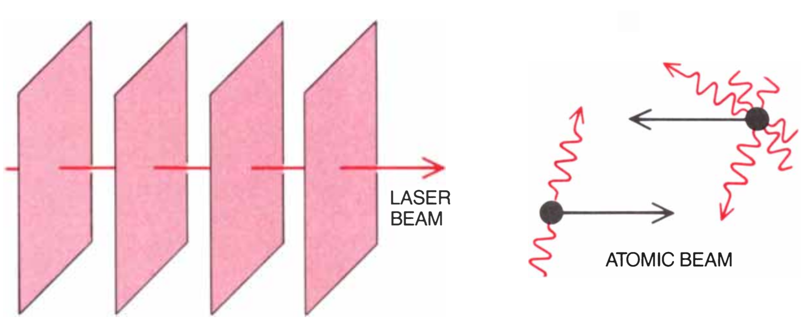
All the approaches begin with freely moving neutral atoms, usually sodium atoms in the form of a beam. We generate the sodium beam by heating an oven containing sodium metal to 450 degrees C. The metal vaporizes, and the atoms come out of the oven through a small pinhole as a diverging beam; we mask off all but a narrow fraction of the beam by positioning another pinhole about 10 centimeters away. In our apparatus a laser shines directly into the sodium beam. In this way each atom can interact with the light for as long as possible.

When an atom absorbs light, it jumps to an excited state. The atom can return to the ground, or unexcited, state by one of two processes, stimulated emission or spontaneous emission. If the emission is stimulated (induced by the laser light), the emitted photon travels in the same direction as the absorbed photon and the atomic momentum does not change. If the emission is spontaneous, however, photons are emitted in random, symmetrically distributed directions. Repeated absorptions followed by spontaneous emissions result in a net deceleration of the atoms in the direction of the laser beam. The maximum deceleration a sodium atom can experience in laser cooling is about a million me-





IDENTICAL ATOMS moving with different velocities in a room-temperature gas appear different to an observer because of a phenomenon known as the Doppler shift (*top*). The basic effect is familiar to anyone who has heard the sudden change in pitch of the horn of a passing train. If the atoms were stationary, they would absorb and emit radiation of almost the same frequency. Atoms moving toward the observer appear to radiate at slightly higher frequencies (*blue*), because the crests and troughs of their radiation waves reach the observer at a higher rate. Atoms moving away from the observer appear to radiate at slightly lower frequencies (*red*). The effect of the Doppler shift is to smear out what would otherwise be a sharply defined spectrum, or distribution of frequencies (*left*). The smeared-out spectrum has the same shape as the Maxwell-Boltzmann distribution in the illustration on the opposite page.



LASER COOLING of a gas takes place when atoms "see" more "head wind" than "tail wind." Such conditions are realized by tuning a laser beam to a frequency lower than the frequency strongly emitted and absorbed by the atoms. Some of the atoms moving toward the laser beam will then have the appropriate Doppler shift to absorb and reradiate the light, and they will slow down. Atoms moving with the laser beam, on the other hand, see the frequency of the light shifted still lower, and they are less likely to absorb the light: they will be speeded up very little. Consequently on the whole the atoms slow down.

ters per second squared, which is close to 100,000 times the acceleration of gravity at the surface of the earth. At such an enormous deceleration a sodium atom with a velocity of 1,000 meters per second would be stopped in one millisecond over 50 centimeters.

As the atoms slow down, even by a few meters per second, their Doppler shift changes enough to inhibit their absorption of light. Eventually the atoms will stop decelerating and will continue their journey unimpeded by the laser beam. One way to compensate for this undesirable effect is to sweep the laser beam to higher frequencies as the atoms slow down so that the atoms continue to absorb the radiation. V. S. Letokhov and his colleagues at the Moscow Institute of Spectroscopy proposed this approach in 1976, and John Prodan and one of us (Phillips) first demonstrated it in 1983 in our Gaithersburg laboratory. Since then John L. Hall and his co-workers at the National Bureau of Standards in Boulder and a number of other groups have successfully employed the approach.

Throughout most of our work, however, we have exploited a different technique to overcome the complication caused by the changing Doppler shifts. We hold the frequency of the laser beam constant and manipulate the energy levels of the atoms so that the atoms continue to absorb the beam. We do this by sending the atoms through a magnetic field whose strength varies along the path traveled by the atoms. The energy levels of an atom change in a well-determined way when the atom is placed in a magnetic field, a phenomenon known as the Zeeman effect. We have tailored the magnetic field in our apparatus so that it is strong at the point where the atoms first enter it and gradually tapers off in strength with increasing distance from the point of entry. As an atom passes through such a magnetic field its energy levels continually change so that they compensate for the decreasing Doppler shifts. As a result the atom continues to absorb light from the laser beam and to decelerate.

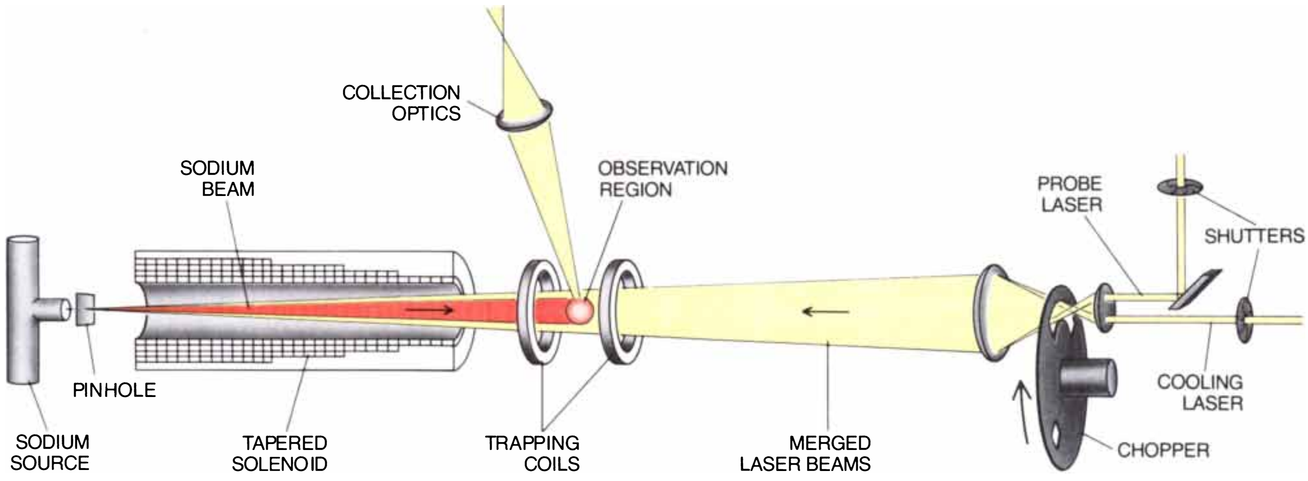
With our technique we have cooled a beam of atoms to less than 100 millikelvin and brought the average velocity of these atoms to zero. In effect we have produced a cold gas of nearly stopped atoms.

Laser-cooling a gas of atoms, in contrast, is quite different from laser-cooling a beam of atoms. The reason is that all atoms in a beam move in approximately the same direction, so that a single laser beam suffices to oppose their motion. On the other hand, free atoms in a gas move in all directions and therefore require several laser beams to cool them. At the AT&T Bell Laboratories, Steven Chu and his colleagues, having laser-cooled a beam of atoms, employed multiple laser beams to further cool the sample to less than a millikelvin. Atoms at the intersection of the multiple beams experience a retarding force in any direction in which they move and are therefore said to be in an "optical molasses."

One of the exciting consequences of stopping neutral atoms is that they can then be held in atom traps: bottles for atoms whose "walls" are electromagnetic fields rather than material substances. It has long been possible to trap electrons, ions and other charged particles, because strong electric and magnetic fields can greatly affect their motion [see "The Isolated Electron," by Philip Ekstrom and David Wineland; SCIENTIFIC AMERICAN, August, 1980]. The same electric and magnetic fields have little effect on a neutral atom, however, since it carries no net charge.

A number of different traps for neutral atoms have been proposed over the past 25 or 30 years. In the 1950's Wolfgang Paul of the University of Bonn suggested that, in principle, magnetic traps could be employed. Another alternative, laser traps, was proposed in about 1970 independently by Letokhov and by Arthur Ashkin of Bell Laboratories. In 1978 Paul and his colleagues succeeded in trapping neutrons in a magnetic storage ring. Their experiment was a landmark one, because it was the first time that neutral particles had been confined electromagnetically. Their work enabled them to make new measurements of the average lifetime of the neutron. (After about 15 minutes an isolated neutron decays into a proton, an electron and a particle called a neutrino.)

Our trap for neutral atoms exploits the same physical principles used in Paul's neutron storage ring. Even though a neutral atom has no net electric charge, it can still have a small magnetic dipole: the atom can behave as if it were a tiny bar magnet. Now, if a bar magnet is immersed in an inhomogeneous magnetic field, the strength of the field at one pole differs



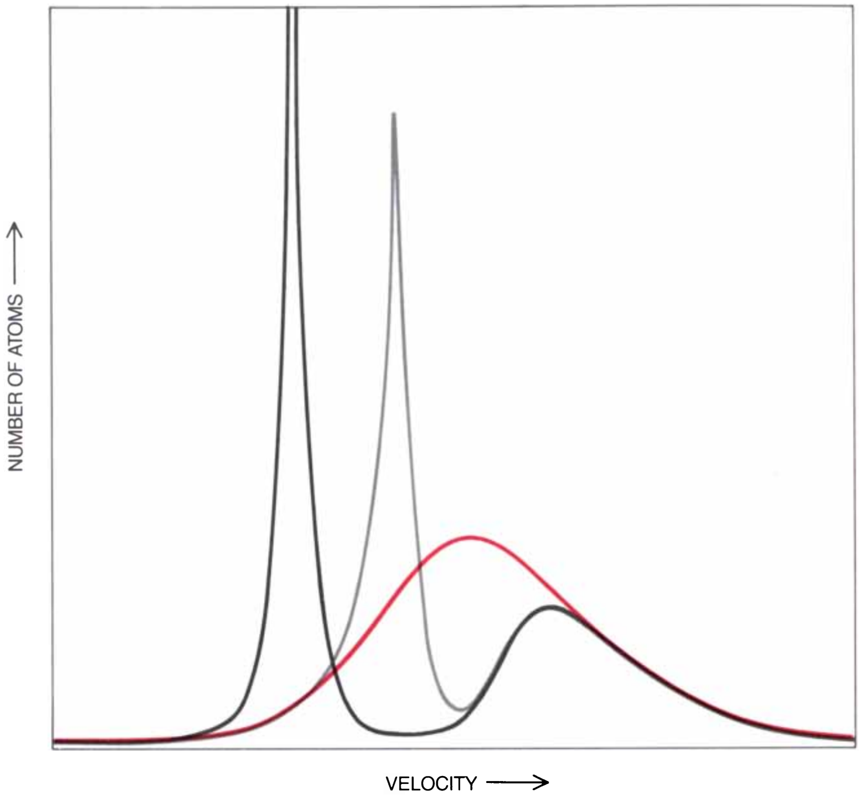
AUTHORS' APPARATUS for cooling an atomic beam and magnetically trapping neutral atoms is shown schematically. The investigators slow a beam of sodium atoms traveling through a solenoid by shining a laser beam on the atomic beam. Measurement of the velocity distribution of the atoms and the laser-induced changes to it is done by collecting and detecting the fluorescence from atoms excited by a second, very weak probe laser propagating nearly parallel to the atomic beam. The absorption and hence the intensity of this fluorescence will depend on the atomic velocity because of the Doppler shift of the probe beam. The dependence of the fluorescence on the probe-laser frequency reflects the velocity distribution of the atoms. Very slow atoms can be trapped in the magnetic field generated by the pair of "trapping coils." The trapped atoms move as if they had a temperature of 10 millikelvin, slightly above absolute zero. The chopper and shutters are used to turn the laser beams on and off at appropriate times.

from the strength at the other pole, and a force will be exerted on the magnet. If the bar magnet happens to be an atom, the force will be quite small, but it is nonetheless an observable quantity. The effects of an inhomogeneous magnetic field on a neutral atom were first demonstrated in 1924 by Otto Stern and Walther Gerlach. The Stern-Gerlach experiment showed that a silver atom can be thought of as a bar magnet whose axis can have only two possible orientations with respect to the magnetic field. (An ordinary bar magnet, of course, can have a continuous range of orientations.)

The sodium atoms in our experiments have two classes of orientations: one class in which the atoms are attracted to strong magnetic fields and one in which they are repelled by strong magnetic fields. In our process of laser cooling we optically pump all the atoms into the orientation repelled by strong magnetic fields. To trap the atoms we therefore constructed a pair of current-carrying coils arranged so that their magnetic fields oppose each other. There is zero magnetic field at the midpoint between them, from which the strength increases in any direction, and so the atoms are pushed toward the midpoint. This kind of trap was one of several early proposals made by Paul; similar traps have been proposed for the confinement of ultracold neutrons [see "Ultracold Neutrons," by R. Golub, W. Mampe, J. M. Pendlebury and P. Ageron; SCIENTIFIC AMERICAN, June, 1979].

In our trapping experiments we decelerate atoms to low velocities by laser-cooling them. We then let the atoms drift into the space between the two current-carrying coils and bring the atoms to rest with a short pulse of light, a technique we developed in our laboratory with Prodan, Alan Migdall, Jean Dalibard and Ivan So. The residual motion of the atoms in the sample is so small—the atoms travel at speeds of only a few meters per second or so—that once the atoms have entered the region between the coils there is ample enough time to turn on the electric current that energizes the magnetic trap.

Our trap has proved successful in catching and confining some of the atoms in the sample, as experiments we have done with Thomas H. Bergeman of Stony Brook and the above collaborators have shown. The major loss of atoms from the trap is due to random collisions with background gas molecules. Since the magnetic force that holds the atoms in place is so very small, the trap is said to be "shallow." An atom held in the trap is a "sitting duck" if it is struck by a room-temperature molecule. Even though the apparatus is placed in a vacuum, there are still enough stray atoms moving around to cause destructive collisions and impart enough energy to eject atoms from the shallow trap.



EFFECTS OF LASER COOLING on the distribution of velocities in a beam of atoms are pronounced. Without laser cooling the velocity distribution is wide (*colored curve*). When the atomic beam is opposed by a laser beam and travels down a long solenoid that produces a uniform magnetic field, the velocity distribution is changed: some atoms near the center of the distribution are slowed and concentrated into a fairly narrow peak where all atoms have roughly the same velocity (*gray curve*). By tailoring the magnetic field so that it is strong at the point where the atoms first enter it and gradually tapers off in strength with increasing distance from the entry point, many more atoms are slowed to a much lower velocity and concentrated into a sharper velocity distribution (*black curve*). It is important to distinguish between deceleration, the reduction of velocity, and cooling, the reduction of velocity spread. The authors' method achieves both.

The maximum speed a sodium atom can have and still be held in our trap is 3.5 meters per second, which corresponds to an energy in temperature units of about 17 millikelvin. We have trapped tens of thousands of atoms with this energy and less for periods longer than one second in a volume of about 20 cubic centimeters, which is roughly the volume of a golf ball. The limitation on the trapping time is due entirely to collisions with any stray atoms left in the vacuum. In a perfect vacuum the ultimate limitation on trapping time would be the rate at which an atom makes a quantum transition from a state that is repelled by strong magnetic fields to a state that is repelled by weak magnetic fields. Calculations show that most atoms would remain trapped for many hours.

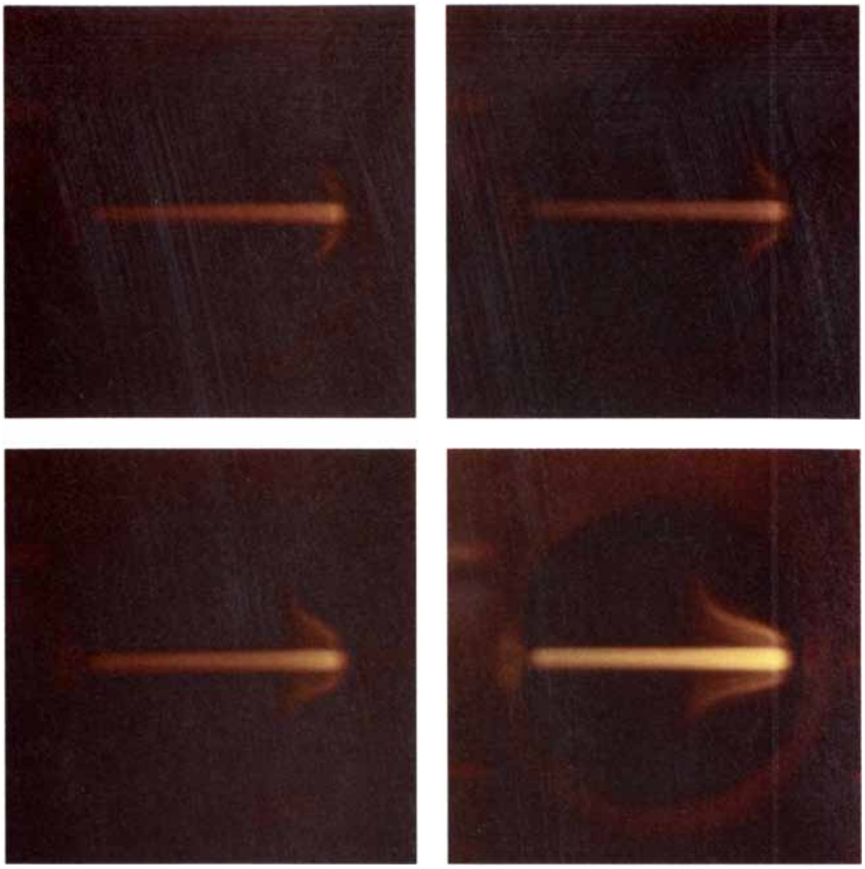
Another method of trapping neutral atoms relies entirely on laser beams and the forces they exert. In one version of such a trap the forces do not come from the simple transfer of momentum that occurs when an atom absorbs a photon and spontaneously emits another photon, as in the laser-cooling process itself. Rather, they come from a subtler and potentially stronger process. The laser beam, which consists in part of an oscillating electric field, induces dynamic changes in the atom that result in a force if the laser field is nonuniform (just as in the case of a magnetic moment in a nonuniform magnetic field). By convention such a force is called a dipole, or gradient, force.

The dipole force can be used to confine atoms, as Letokhov suggested in 1968. In 1978 Ashkin proposed a particularly simple and elegant form of laser trap, which has just recently been demonstrated. In Ashkin's design a single laser beam is focused to a small spot. The focusing produces a laser field that is strongest at the center of the focus; the field decreases with increasing distance in all directions away from the focus. When the laser frequency is tuned below a frequency at which the atoms absorb strongly, the dipole force pulls the atoms into the strongest part of the field. Of course, the ordinary radiation pressure resulting from momentum transfer tends to push the atoms along the direction in which the laser beam travels, but the dipole force can overcome the unwanted effect if the laser beam is tightly focused, is intense enough and is tuned far enough away from the absorbing frequency of the atoms. Still, the depth of a laser trap is very small, and Ashkin's design had to wait, along with the magnetic trap, for the success of laser-cooling techniques before it could be implemented.

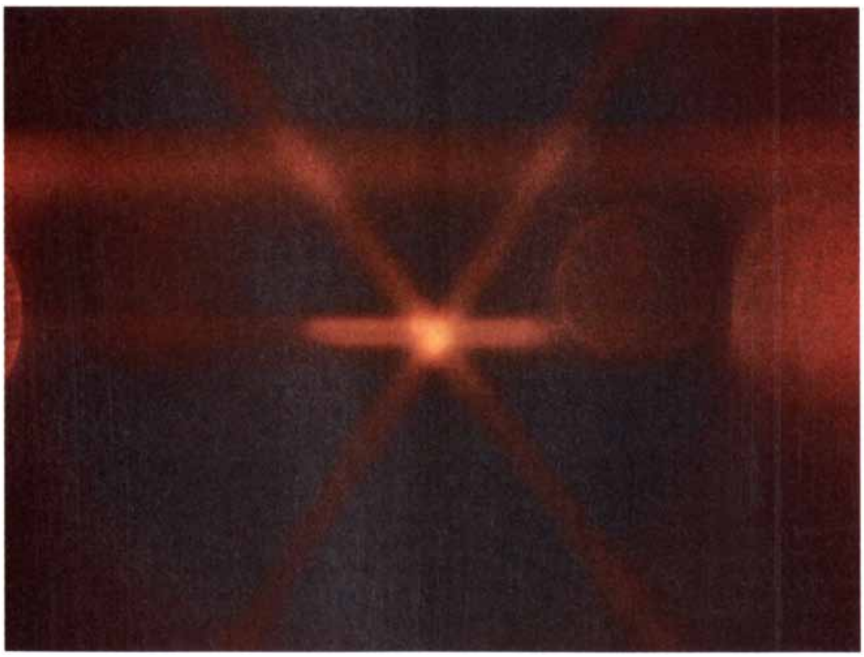
An additional complication of laser traps is that, in contrast to magnetic traps, the trapping field itself tends to heat the atoms so that they "boil" out of the trap. The solution to the problem is to continue to laser-cool the atoms while they are in the trap. The precise scheme for doing this, which involves switching rapidly between cooling and trapping by turning various laser beams on and off, was proposed by Dalibard, Serge Reynaud and Claude Cohen-Tannoudji of the École Normale Supérieure. All these techniques were recently combined, along with three-dimensional cooling in optical molasses, by Chu and his colleagues to make a tiny laser trap for sodium atoms in which a few hundred atoms are held in a volume of 10^{-7} cubic centimeter at a temperature below a millikelvin.

Even more recently a group of investigators in Cohen-Tannoudji's laboratory demonstrated that an atomic beam can be cooled by means of dipole forces. Their method shows great promise, because the strength of the dipole force allows the cooling to proceed much faster than it can with ordinary laser cooling.

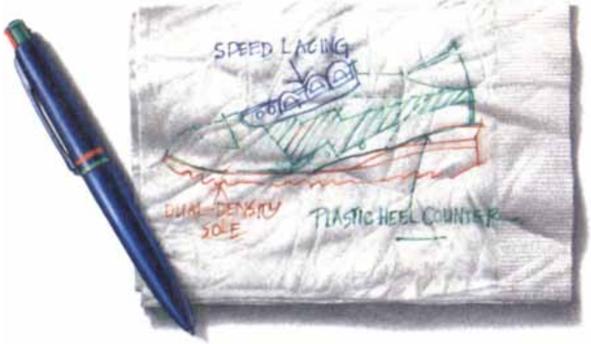
Although the future of laser cooling and trapping of neutral atoms and ions is hard to predict, it is clear they are solidly established as major areas of research. Thermal motion, which has long bedeviled investigators, is in rapid retreat. Atomic beams and gases that have millikelvin temperatures are now easily generated; microkelvin temperatures and below may soon be possible. A new era in atomic measurement is at hand.



LASER-COOLED BEAM of sodium atoms is shown as it emerges from the solenoid in the authors' apparatus. The atomic beam travels from right to left; the laser beam moves from left to right. As the frequency of the laser is decreased, the position at which the sodium atoms stop moves to the left and more of the skirt comes into view (*left to right, top to bottom*). The beam is viewed through a circular port in the side of the apparatus.



OPTICAL MOLASSES produced in the authors' laboratory by Phillip Gould and Paul Lett is seen as a bright spot at the intersection of six laser beams. The six laser beams strongly oppose and quickly damp any atomic motion in the intersection region, as if the atoms were in molasses. Laser-cooled sodium atoms enter the molasses from the left and get stuck. A cooling laser beam illuminates some of the atoms (*top horizontal streak*).

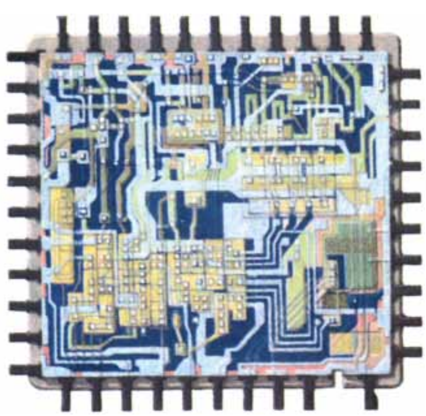


Who can help this design get off on the right foot?

Who offers software that can help the concept pick up speed?

Whose cellular network in New York can the CEO use to okay the project?

Who provides inventory control software that can keep track of manufacturing?



Whose digital switching system can the project team use to communicate with each other?

Who makes it easier for good ideas to get off and running?

THE ANSWER IS NYNEX

Telecommunications • Information Systems • Software • Publishing • Business Services

For information, 1-800-535-1535

SCIENCE AND THE CITIZEN

Hobbling High-Tech

High-technology industries in the U.S. have long complained about the maze of procedures they must negotiate in order to obtain licenses to export "dual use" products—those having military as well as civilian applications. Their competitors in other countries (even in U.S.-allied countries) are subject to much less stringent export controls, and so they gain an advantage in selling such products as high-density semiconductor chips, automated machine tools and precision measuring instruments. At the same time the U.S. provokes resentment by attempting to impose legal obligations, such as guarantees that products will not be reexported, on recipients of the items. Customers are further deterred, critics say, by complex U.S. regulations that create delays and uncertainty in licensing.

Critics both at home and abroad recently received some influential support from a two-year study conducted by a special panel of the National Academy of Sciences and the National Academy of Engineering. The panel was chaired by Lew Allen, Jr., a former director of the National Security Agency and now director of the Jet Propulsion Laboratory. The list of other members includes Melvin R. Laird, Secretary of Defense under President Nixon, Admiral Bobby R. Inman, a former deputy director of the C.I.A. who now heads a fledgling industrial electronics enterprise, Joshua Lederberg, president of Rockefeller University, and G. William Miller, an industrialist who headed Textron, Inc., and served as Secretary of the Treasury under President Carter.

The impressive credentials of the panel did not keep the Department of Defense from disparaging the study. Assistant Secretary of Defense for international security policy Richard N. Perle dismisses the report as the unproved work of interested parties from the business community. The Pentagon stopped cooperating with the panel in February of last year and withheld the second of two promised $100,000 support payments. According to a Defense Department official, that was because the study "was not producing new information or practical recommendations."

No one disputes the need to keep critical technologies out of the hands of the Soviets, the Allen panel says; fearsomely complex U.S. regulations, however, do not succeed very well because they attempt to cover far too many technologies, including a large number widely available on international markets. At the same time the regulations could be depriving U.S. industry of about $9 billion in exports—and U.S. labor of 200,000 jobs—per year. Furthermore, the panel argues, the military strength of the U.S. and the North Atlantic Treaty Organization depends entirely on a technology lead, which could be eroded by handicaps that burden manufacturers.

The Allen panel points out that because the Defense, Commerce and State departments have together failed to institute a working mechanism for decontrolling ubiquitous technologies, U.S. companies face difficulties in exporting even commonplace products. Some manufacturers have had trouble selling abroad products containing the 8086 microprocessor—which, since it is at the heart of many personal computers, is available the world over. A member of the study's technical staff reports that in Singapore panel members saw Soviet merchant ships anchored at a public quay alongside a street of shops that sell personal computers; staff members found chips that are supposed to be controlled moving over the counter in Japan.

To police the availability of sensitive technology the Allen panel recommends that the U.S. strengthen the international control regime administered by the Paris-based Coordinating Committee on Multilateral Export Controls (CoCom). CoCom consists of the NATO countries (except Iceland) and Japan. It maintains a list of hundreds of technologies that require review to determine whether they can be exported to the Soviet bloc.

The U.S., the panel says, should also concentrate on effective control of a small number of items that are genuinely critical to security. The U.S. should not, except in some instances, impose unilateral controls beyond the ones exercised by CoCom, the panel says. Attempts should be made to get nations outside CoCom to fall in with its procedures; in exchange the U.S. would drop its reexport regulations. Allen also expresses concern about several recent moves by the Pentagon that tighten control of unclassified technical information from Defense-funded research projects.

In addition to criticizing Defense, Commerce and State, the panel says that U.S. export-control policy lacks "affirmative leadership at the highest level." To remedy the problem the panel recommends that the National Security Council be given a stronger role in patching up the interagency quarrels that have led to the present disarray. A far-reaching review by the National Security Council indicates that Allen's study is being taken seriously in the Executive Branch, and there has been interest on Capitol Hill, where major trade legislation will be considered during the year.

SDI Goes on the Offensive

Two years after it was launched as a $26-billion research project the Strategic Defense Initiative (SDI) faces a potentially fractious Congress, opposition from influential parts of the scientific and arms-control communities, and the certain end of the Reagan Administration. To cope with this dubious future Secretary of Defense Caspar W. Weinberger has followed the old football maxim that the best defense is a good offense. He announced that the SDI might be taken out of the laboratory and actually deployed as a "phase 1" system in the early 1990's.

To meet this deadline the system would foresake for the time being laser and particle-beam technology. Instead it would employ chemical rockets, some launched from satellites, that would intercept incoming warheads and destroy them on impact. The system could not destroy all incoming missiles. Its existence would, however, ensure that missile defenses would be "in place and not tampered with by future administrations," as Attorney General Edwin Meese III has put it.

A decision for early deployment would also mean the U.S. must abandon the 1972 Anti-Ballistic Missile (ABM) Treaty within the next few years. The system would include dispersed ground- and space-based components embodying conventional ABM technologies. Spurgeon M. Keeny, Jr., President of the Arms Control Association, says early deployment would mean "deferring arms control progress for the foreseeable future."

The idea of a defense based on rocket interceptors actually predates the SDI: The "High Frontier" organization of retired Air Force Lt. Gen. Daniel O. Graham made such a proposal in 1981, and his vision may have influenced President Reagan to launch the SDI. The High Frontier proposal was found lacking by most analysts, chiefly because its orbiting battle stations would be vulnerable to an attack by ground-launched rockets or by space

As you see, the exterior speaks for itself. But the interior asked us to say a few words. *Words like "redesigned analog instrumentation," and available "articulating bucket seats" describe Grand Am SE's new interior. Add quick handling, V6 power and its awesome exterior to this inside information and you have the true meaning of another word: "Excitement."*

PONTIAC GRAND AM

WE BUILD EXCITEMENT

mines. Yet a similar idea was proposed again last December by the George C. Marshall Institute, a private research group based in Washington. The institute is home for some well-known proponents of missile defense, including John L. Gardner, former director of systems for the SDI Organization; the new Pentagon view of early deployment is based on a scheme similar to the Marshall Institute's proposal.

The plan envisions a three-layered defense. Space-based kinetic-kill vehicles—11,000 of them carried on several thousand small satellites—would attack Soviet missiles during the boost and postboost phases. Ten thousand ERIS (Exoatmospheric Re-entry Vehicle Interceptor Subsystem) interceptors launched from the ground would tackle warheads in midcourse; ERIS, a descendant of the Homing Overlay Experiment, is considered the maturest of the technologies. Finally, 3,000 HEDI (High Endoatmospheric Defense Interceptor) interceptors would mop up warheads that made it through the first two layers. Tracking would be done by 14 sensor satellites, as well as by ground- and air-based systems.

The Marshall Institute contends its system would destroy at least 93 percent of incoming warheads, more if progress is made on the difficult problem of distinguishing decoys from real warheads in midcourse. If a decision were made in 1987, the institute says that deployment of a system costing $121 billion could start by 1994.

Those claims are disputed by at least two prominent defense analysts. Ashton B. Carter of Harvard University thinks the scheme "glosses over the hard parts." The Marshall Institute's account of how satellites are to survive attack from the ground is "too vague to be believed," Carter says, and it ignores the threat of space mines. Carter is also not convinced that the critical boost-phase layer is cost-effective at the margin, a criterion for deployment that has widespread support in Congress. Richard L. Garwin of IBM's Thomas J. Watson Research Center argues that the institute's estimates of cost are too low and that effectiveness is exaggerated. Garwin points out that critical countermeasures, such as fast-burn boosters carrying many decoys, are not considered.

Why is the Pentagon suddenly considering early deployment of a relatively simple kinetic-kill system looking much like schemes already examined and rejected? The plan is certain to meet stiff resistance on Capitol Hill, where lawmakers have insisted that SDI funds go only to research. The answer may be that SDI advocates have decided that the certainty of having a system with deployed components—and hence a political constituency—within 10 years is worth any storm over abandoning the ABM Treaty.

Veteran U.S.-Soviet-relations expert Marshall D. Shulman believes the attention to near-term deployment is designed to force the president to abandon his proposal made to the Soviet Union to honor the ABM Treaty for the next 10 years. Although the SDI budget has jumped from $1.4 billion in 1985 to $3.53 billion this year, the Department of Defense wanted much more. The prospect of even an imperfect shield against nuclear missiles, rather than just more research, might be politically unstoppable.

The Irrational Stock Market

On January 23 the Dow Jones industrial average surged ahead 64 points and then plummeted 114, a more precipitous zigzag than on any previous day in the stock market's history. Asked by reporters to explain this volatility, some Wall Street watchers blamed computerized selling programs, profit taking or the falling dollar. Other observers decried the day's events as "absolutely absurd," "crazy" and "irrational."

Exactly so. The market has always been somewhat irrational, according to Robert J. Shiller of Yale University. Based on his reading of more than a century's worth of economic data, Shiller maintains that the prices of stocks have often soared too high or sunk too low for no good economic reason. This conclusion flies in the face of one of the reigning theories of modern economics: the efficient-markets model. Proponents hold that the prices of stocks at any given moment reflect perfectly all publicly available information about their future value. The model gives a flattering picture of the investment community—at least in its sum—as an omniscient investor who instantaneously digests information and buys or sells stocks accordingly.

A corollary of this view is the random-walk theory, which posits that trying to find stocks that are "good buys" is futile; the omniscient investor has already bought them and pushed their price up. Because tomorrow's prices will change only with tomorrow's news—by definition unpredictable—any change will be random.

Shiller contends that if the stock market were truly efficient, the fluctuations in prices should roughly parallel fluctuations in indicators of stocks' true investment value. He tested this hypothesis with three generally accepted indicators: dividends, which provide income and reflect a company's attitude toward its own future; interest rates, which affect the relative value of stocks compared with no-risk, interest-bearing investments; and per capita consumption, which, as it rises or falls, encourages buying or selling of stocks, driving prices up or down.

After plotting these indicators for the years from 1871 through 1986, Shiller found that during many periods—such as the Depression and the early 1970's—stock prices fell lower or rose higher than the efficient-markets model would predict. Writing in *Science,* he concludes that "most of the volatility of stock market prices appears unexplained."

Shiller suggests that the omniscient investor efficient-markets theorists describe may sometimes overreact to financial news or succumb to "social optimism or pessimism, fashions or fads." An astute individual, Shiller says, can take advantage of this "foolishness" by buying when stocks are too low or selling when they are too high. "I think an intelligent person can do better than random," he observes. Shiller does not play the market himself, however. "I may in the future," he says. "I'm not a rich man."

Burton G. Malkiel, dean of the Yale School of Organization and Management and author of *A Random Walk Down Wall Street,* disagrees. Stocks may indeed prove to have been priced too high or low, Malkiel concedes, but only in retrospect.

In the early 1970's, for instance, when Watergate and Vietnam dominated headlines, the stock market plunged because, Malkiel recollects, "people thought the world was going to hell in a hand basket. It turns out that they were wrong. But were they stupid? Was the market inefficiently priced, given what people knew then?" Absolutely not, Malkiel declares, and adds: "I defy Bob [Shiller] to tell you whether the market is too high or too low now."

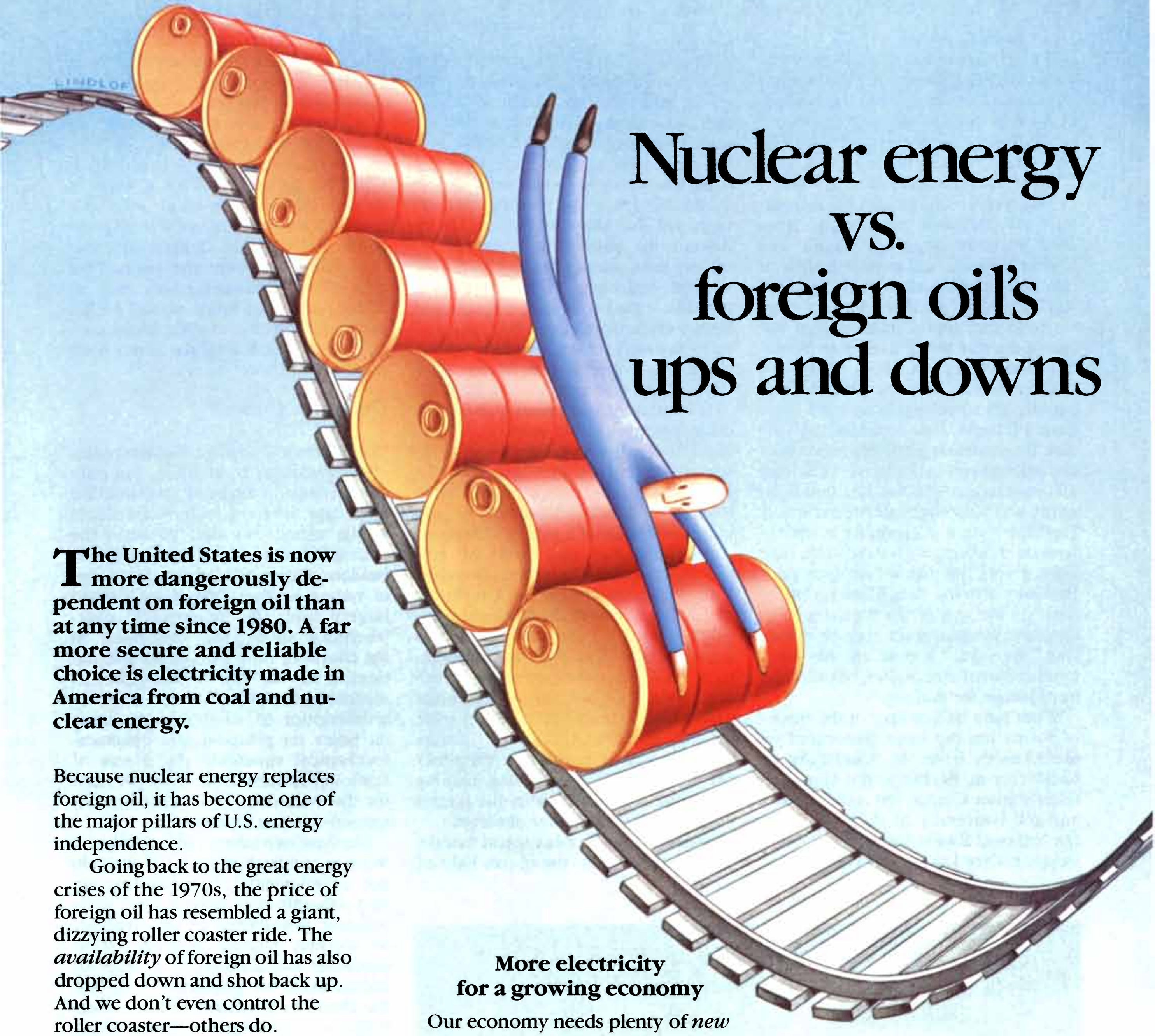
Shiller replies: "I think it's right in the middle."

PHYSICAL SCIENCES

Birth Announcements

The 169th annual meeting of the American Astronomical Society heard reports of what may be the birth of a planetary system and the birth of a galaxy. The discovery of an entirely new kind of astronomical object was also indicated.

"The strongest evidence yet" that a nearby star may be surrounded by a planetary or protoplanetary system was reported by Francesco Paresce



Nuclear energy vs. foreign oil's ups and downs

The United States is now more dangerously dependent on foreign oil than at any time since 1980. A far more secure and reliable choice is electricity made in America from coal and nuclear energy.

Because nuclear energy replaces foreign oil, it has become one of the major pillars of U.S. energy independence.

Going back to the great energy crises of the 1970s, the price of foreign oil has resembled a giant, dizzying roller coaster ride. The *availability* of foreign oil has also dropped down and shot back up. And we don't even control the roller coaster—others do.

Consider this ominous statistic: even though prices were low in 1986, America still had to pay about $30 billion for foreign oil. That's a lot of dollars leaving this country, adding to an already huge trade deficit.

Nuclear energy cuts oil imports

Clearly, the more energy we use in the form of electricity from coal and nuclear energy, the less oil we have to import.

Nuclear-generated electricity has already saved America over two billion barrels of oil, with billions more to be saved before the turn of the century. That's why it's so important for our energy self-reliance.

More electricity for a growing economy

Our economy needs plenty of *new* electrical energy to keep on growing. Almost all of that new energy is coming from coal and nuclear electric plants.

The truth is that nuclear energy is an everyday fact of life in the U.S. It's been generating electricity here for nearly 30 years. Throughout the country are more than 100 nuclear plants, and they are our second largest source of electric power. As our economy grows, we'll need more of those plants to avoid even more dependence on foreign oil.

Safe energy for a secure future

Most important, nuclear energy is a safe, clean way to generate electricity. U.S. nuclear plants have a whole series of multiple backup safety systems to prevent accidents. Plus superthick containment buildings designed to protect the public even if something goes wrong. (It's a "Safety in Depth" system.)

The simple fact is this: America's energy independence depends in part on America's nuclear energy.

For more information, write the U.S. Committee for Energy Awareness, P.O. Box 1537 (RC20), Ridgely, MD 21681. Please allow 4-6 weeks for delivery.

Information about energy America can count on

U.S. COMMITTEE FOR ENERGY AWARENESS

and Christopher Burrows, astrophysicists with the European Space Agency who are on assignment to the National Aeronautics and Space Administration's Space Telescope Science Institute. The star is Beta Pictoris, a mere 53 light-years from the solar system.

Two years ago, by analyzing near-infrared radiation emanating from Beta Pictoris, Bradford Smith and Richard Terrile discovered a disk of relatively cool matter encircling the star. Pressing further, Paresce and Burrows gathered a fuller optical image of the star at the European Southern Observatory in La Silla, Chile.

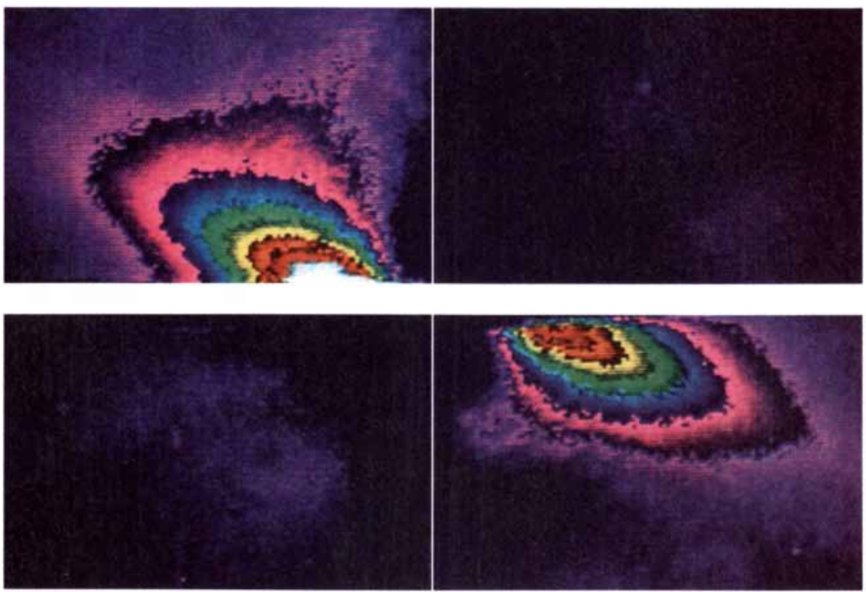
They found that the disk reflects equally all wavelengths of light from Beta Pictoris. This suggests that the disk incorporates particles more than one micrometer in diameter, or at least 10 times larger than the dust that ordinarily wafts through interstellar space. Dust less than a micrometer in diameter would reflect short-wavelenth, blue light, giving the disk a blue-gray cast. Burrows says the data place no upper limit on the size of the particles. "An agglomeration process may be occurring," he says. "It may already have created planet-size bodies, but we have no evidence for that yet."

What may be a galaxy in the throes of formation has been discovered by astronomers from the University of California at Berkeley, the Harvard-Smithsonian Center for Astrophysics and the University of Arizona. With the National Radio Astronomy Observatory's Very Large Array and several optical telescopes the group located a luminous object three times the size of the Milky Way and about 12 billion light-years away, placing it among the most distant objects in the visible universe in both space and time.

By analyzing light from the object, called 3C 326.1, the investigators determined that the vast bulk of its radiation—the equivalent of about 10 billion suns—comes from a cloud of ionized hydrogen; a much smaller amount, equal to about 100 million suns, comes from stars. This lopsided ratio, the workers say, suggests that the cloud is a "young and/or forming galaxy." They note, however, that 3C 326.1 emits the enormous amounts of radio energy usually associated with black holes. It was these powerful radio waves, in fact, that first drew the attention of the investigators. Most theorists believe black holes form only in the latter part of a galaxy's lifetime.

An altogether new kind of cosmic structure may have been found by Vahe Petrosian of Stanford University and Roger Lynds of the National Optical Astronomy Observatories. Working at the Kitt Peak National Observatory, they discovered three vast luminous arches partially encircling a large galaxy at the center of three separate galactic clusters. One arch appears to be broken into sections; the other two appear to be continuous, making them, the workers believe, the largest luminous structures ever observed.

Petrosian and Lynds suggest that the arches, which emit the bluish light of relatively young stars, might represent the edge of a shock wave traveling outward from an explosion with the force of millions of supernovas; another theory is that the arches consist of stars dragged outward from a galaxy by a powerful gravitational force, such as another galaxy or a black hole. Although the arches are in different parts of the sky, they are all about five billion light-years from the earth. This coincidence, Petrosian offers, may indicate that "these things existed until a particular epoch and then broke up." He adds: "I think we'll see others if we look farther away."



DISK OF PARTICLES surrounds Beta Pictoris, a star 53 light-years from the earth. To obtain this image astronomers at the European Southern Observatory electronically subtracted an image of a "reference star" that resembles Beta Pictoris from the image of Beta Pictoris itself. White lines indicate where the image was condensed. False colors correspond to different intensities of light. The faintest light is violet; the strongest, red.

Quantum Chaos?

Tomorrow's weather, like the radioactive decay of an atom, can only be described in terms of probabilities. Two very different factors contribute to the uncertainty that obscures the course of these events. The chaotic, or random, nature of such a macroscopic system as the atmosphere derives largely from the innumerable overwhelming effects that can result in the course of time from small fluctuations in the variables of deterministic, mechanical equations. In contrast, a description of subatomic phenomena relies on probabilistic, quantum-mechanical equations; the source of randomness here is an inherent limit on the precision with which certain variables can be measured.

Do these two sources of randomness share a common ground? If they do not, a fundamental, discomforting gap may separate microscopic from macroscopic phenomena in the structure of modern physics. Indeed, recent computer calculations by a team of Italian and Soviet physicists imply that the chaos so prevalent in the macroscopic world is absent in the quantum-mechanical domain. Their work raises the question of how well the theory of quantum mechanics meshes with the theory of chaos; it even challenges the compatibility of quantum mechanics with the physics of the universe in general, according to some physicists.

Good testing grounds for observing how the two theories fit together (or do not) are "semiclassical" systems: interactions of particles and fields that can be treated both classically and quantum mechanically. One such system, the ionization of a highly excited hydrogen atom by microwaves, has been analyzed by Giulio Casati of the University of Milan, Boris V. Chirikov and Dima L. Shepelyansky of the Institute of Nuclear Physics in Novosibirsk, and Italo Guarneri of the University of Pavia. It is generally accepted that a hydrogen atom normally

ionizes whenever the atom's one electron absorbs a single photon whose energy is greater than a certain threshold. What drew the attention of the investigators was experimental evidence that ionization could take place at frequencies, or energies, much below the single-photon ionization threshold.

As the workers report in *Physical Review Letters,* the problem can be approached classically by analyzing the ionization as a chaotic process. Alternatively, the problem can be stated in quantum-mechanical terms. The investigators implemented both theoretical approaches and discovered that under certain conditions both indeed yielded a large probability of ionization at frequencies below the single-photon threshold. In the quantum-mechanical treatment, however, the probability of ionization drops abruptly to zero at a frequency (still below the single-photon threshold) at which the classical probability remains greater than zero.

These results, as Joseph Ford of the Georgia Institute of Technology points out in *Nature,* are interesting on three counts. First, they theoretically verify that ionization can occur below the single-photon threshold. In fact, the calculations of Casati and his colleagues show that under suitable conditions this type of ionization is much more likely to take place than the single-photon ionization. Ford speculates that such an effect, if it actually exists, could be applied to improve the efficiency of electronic devices that currently rely on the single-photon effect.

Second, the range of frequencies in which the classical electron can ionize but the quantum electron cannot provides an opportunity to test the validity of the two theoretical approaches. If precise laboratory measurements indicate that ionization can occur within this range, then quantum-mechanical equations may have to be revised. Finally, the system adds another paradox to the already paradoxical subatomic world: an electron can be described as being either chaotic or nonchaotic, depending on whether one views the system classically or quantum mechanically.

Other physicists agree with Ford that the first two points should be settled by subjecting the theoretical prediction to empirical observation; the paradox, they argue, could be resolved by merely clarifying semantics. Norman H. Packard of the University of Illinois at Urbana-Champaign points out that the classical definition of chaos has to do with the way a system evolves in time; he notes, however, that "time evolution is impossible to follow through quantum-mechanical measurements." Although the time-dependent equations that describe a quantum-mechanical system are well defined, the corresponding physical state of the system is not; the equations define the system as essentially a superimposition of all possible physical states. What is probed when one makes a measurement of a quantum-mechanical system is one of the many possible states that has fortuitously frozen into reality.

Ford, however, maintains that the paradox is irreconcilable. If experimental evidence confirms his opinion, he foresees that the quantum-mechanical view may have to be rejected in favor of a universal theory of chaos.

Superconductors Heat Up

A report in a German physics journal has triggered a cascade of advances in superconductors, materials that at very low temperatures carry electricity with no resistance. For more than a decade investigators have searched in vain for materials that superconduct at temperatures above a chilly 23 degrees Kelvin (−250 degrees Celsius). Now they have found superconductors that work at temperatures more than twice that high, and they are approaching a goal that a year ago seemed unattainable: superconducting at 77 degrees K., the boiling point of nitrogen.

"That's the holy grail right now," Robert C. Dynes of the AT&T Bell Laboratories in Murray Hill, N.J., observes. Achievement of this goal, researchers say, could reduce by orders of magnitude the cost of power generators and transmission lines, magnetically propelled trains, particle accelerators and other technologies.

The breakthrough came last October when K. A. Müller and J. G. Bednorz of the IBM Zurich Research Laboratory reported in *Zeitschrift für Physik* that a mixture of lanthanum, barium and copper oxide began to lose resistance to electricity at 30 degrees K. Most workers had focused on alloys of two metals, such as niobium-germanium, the previous record holder among high-temperature superconductors. Asked why he turned just over a year ago to metal oxides, which had been thought to be unpromising, Müller recalls: "I convinced myself we needed a new concept."

Neither Müller nor any other investigator has determined exactly why the new materials superconduct at such high temperatures. In a paper in *Physical Review Letters* the AT&T workers speculate that "ordinary electron-phonon interaction" is responsible. Phonons are thermal vibrations in the crystal lattice of a material that, when they are strongly coupled with electrons at low temperatures, can induce superconductivity. Paul C. W. Chu of the University of Houston, however, suggests that "a whole new physics may be involved."

The IBM paper inspired activity at laboratories around the world. Within weeks Chu and his colleagues discovered that under high pressure the mixture could superconduct at over 40 degrees K. Investigators at Bell Laboratories reported that a mixture in which barium was replaced with strontium, which has a smaller atomic radius, would superconduct at 37.5 degrees under normal atmospheric pressure. Other organizations reporting similar results include the Institute of Physics of the Chinese Academy of Sciences, the Tokyo Institute of Technology, the University of Alabama and the Argonne National Laboratory.

Workers in the field caution that equating a sharp drop in resistance with superconductivity can lead to exaggerated reports, such as an early one from China (since retracted) of superconducting at 70 degrees K. The problem, Müller explains, is that a substance may abruptly lose much of its resistance at one temperature but may not truly superconduct until it is cooled another 10 degrees or more. A more reliable sign of superconductivity is the Meissner effect: when a substance superconducts, it cannot be magnetized by an externally applied magnetic field. Chu says the Meissner-effect test is difficult to carry out with pressurized samples; he relied on resistance tests to confirm his recent finding that a barium mixture superconducts under pressure at 52.2 degrees.

Nevertheless, Chu says, recent advances "raise hope" that liquid helium, which boils at four degrees K. and is now the only superconductor coolant in use, might be replaced with nitrogen, which constitutes about three-fourths of the earth's atmosphere. Helium could be replaced now with liquid hydrogen (boiling point, 20.3 degrees K.) and neon (boiling point, 27.1 degrees K.), but hydrogen is combustible and considered unsafe, and neon is expensive. Nitrogen, with the next-coldest boiling point, is safe and cheap; it costs about as much as milk, whereas liquid helium costs as much as a fine brandy. With its high boiling point, nitrogen also evaporates more slowly than helium, hydrogen or neon.

Dynes estimates that a cooling system based on nitrogen could be from 100 to 10,000 times cheaper than a helium-based system. He notes, however, that finding a material that can superconduct above 77 degrees K. will not be enough. The metal oxides being tested now are "very brittle," he notes, and more extensive research would be needed to make them suitable for transmission lines and magnets. Müller says substitution of other oxides for copper oxide and other metals for lanthanum, barium and strontium should result in steady improvement.

BIOLOGY

Political Animals

The red wolf vanished from the forests of America some 10 years ago, a victim of the destruction of its habitats and of crossbreeding with coyotes. Fortunately the U.S. Department of the Interior's Fish and Wildlife Service had caught a few individuals and bred them in captivity. This spring, in what will be the first time that a species extinct in the wild is reintroduced into natural habitats, the service will release three pairs of red wolves in the Alligator River National Wildlife Refuge in North Carolina.

Although the red wolf's prospects have brightened, the future of the Federal legislation funding its recovery, the Endangered Species Act, is uncertain. Passed in 1973, the statute protects endangered species—animals and plants in the U.S. and elsewhere—against hunting, trading and any other activities that might harm them or their habitats. The act also charged the Fish and Wildlife Service and, to a lesser extent, the Department of Commerce with helping species to recover.

The act's official term expired in 1985. Congress has given it a temporary extension, but it remains in legislative limbo. Even if Congress formally reauthorizes the act this year (an attempt last year having failed), some observers worry that it will not receive critically needed increases in funding. For more than 10 years the annual budget for endangered-species programs has remained roughly constant: the Fish and Wildlife Service has been allocated about $25 million and the Commerce Department's National Marine Fisheries Service about $5 million. The consequent erosion in spending power, according to Michael J. Bean of the Environmental Defense Fund, has reduced the act's role to slowing the decline of species rather than spurring their recovery.

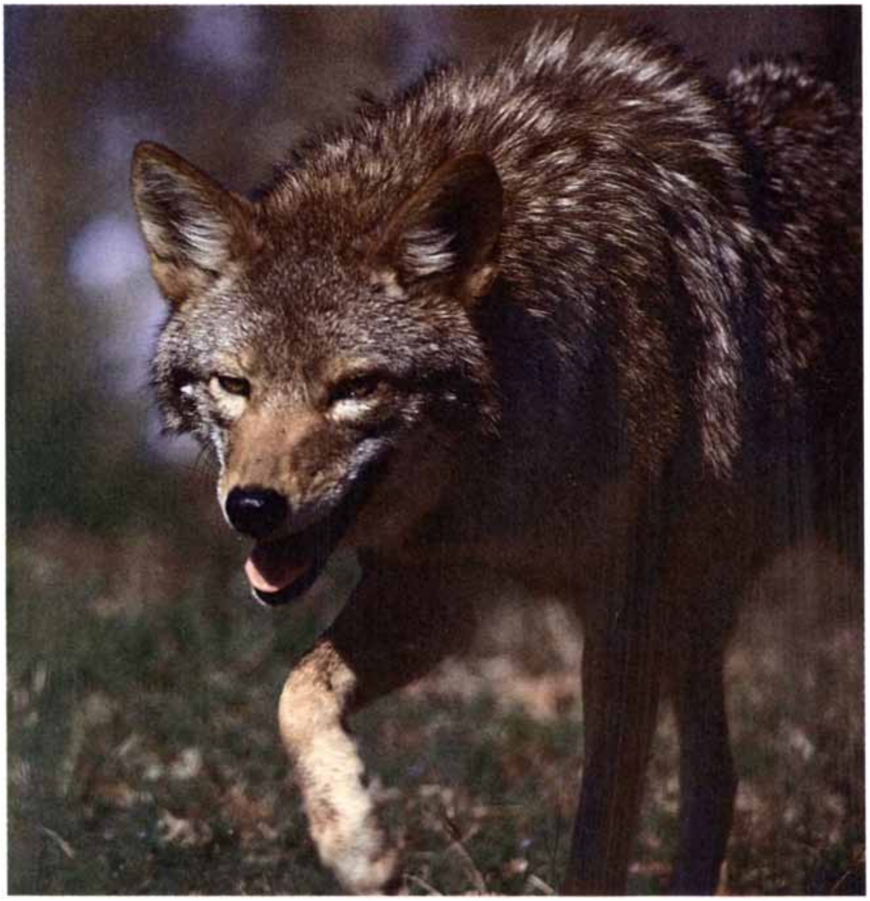
There are some success stories. In recent years the numbers of peregrine falcons, bald eagles, brown pelicans, alligators and grizzly bears, among other species, have increased in certain regions. Bean points out, however, that none of these animals has recovered fully enough to be taken off the list of endangered species. In the entire term of the act only three species, all birds native to small islands in the Pacific Ocean, have earned that distinction.

That leaves 928 species listed as of early this year: 314 mammals, 223 birds, 142 plants, 100 reptiles, 81 fishes, 25 clams, 16 amphibians, 13 insects, nine snails and five crustaceans. J. Scott Feierabend of the National Wildlife Federation notes that the Fish and Wildlife Service has developed recovery plans for many species but lacks the resources to carry them out. "The whole point of the act," Feierabend says, "is to get plants and animals off the list."

Getting on the list is also difficult. There are about 4,000 candidates for listing, according to David C. Klinger of the Fish and Wildlife Service. Before a species—nominated for the list by scientists, environmental groups or the Fish and Wildlife Service itself—is deemed officially endangered, the service conducts an analysis that Klinger calls "inherently conservative." By the time a species is listed it may be too late: the Wyoming toad, listed in 1984, is now thought to be extinct; so are Henslow's sparrow and the Guam bridled white-eye, also recently listed.

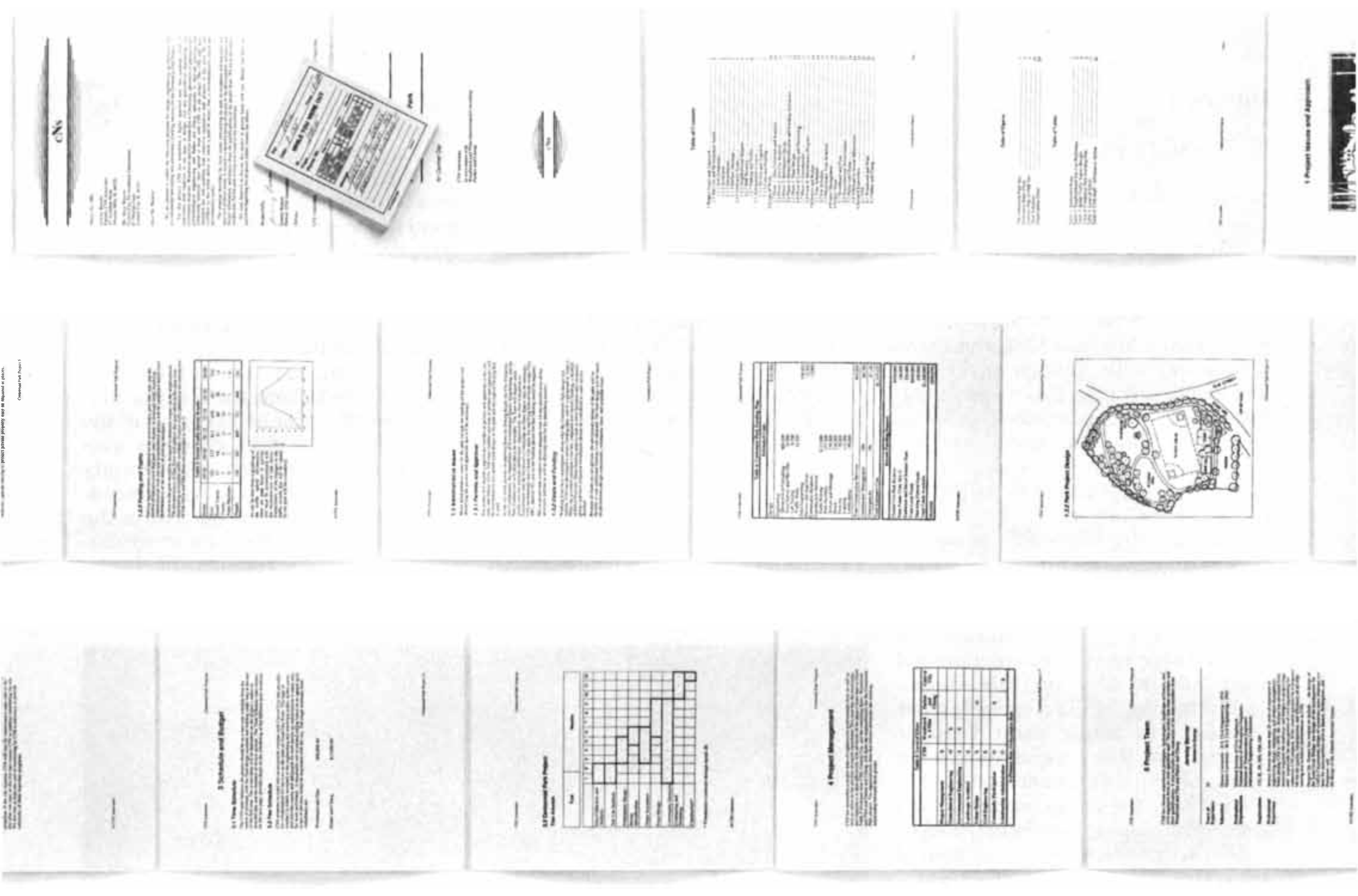
Conservationists would also like better protection of species outside the U.S. The act prohibits the importation of products, such as fur coats, made from endangered species; moreover, it authorizes the Commerce Department to help other countries protect designated species. "The first provision," Feierabend says, "is limited by its nature; the second is limited by lack of funds." Four years ago the Government further weakened protection of nonindigenous species when lawyers of the Department of the Interior ruled that U.S. agencies—such as the Department of Defense or the Agency for International Development—involved in construction projects overseas were not required to ask the Fish and Wildlife Service about the potential impact on local wildlife.

Indeed, the act has not received much support from the Reagan Administration. Almost every year the White House budget proposes cuts in funding for the act; the recently released 1988 budget recommended a cut of almost 20 percent in the Fish and Wildlife Service's programs. Congress usually restores the funds, but its support is by no means unani-



RED WOLF, one of more than 900 species legally protected by the Endangered Species Act, survives now only in captivity. Six wolves will soon be reintroduced into the wild.

How to process docum



You may be content with your present word processor. And it's easy to understand why.

Even the clumsiest word processor is light years beyond the electric typewriter, the accepted standard only a decade ago. Plus, just the term word processing conjures up a rather modest expectation, ho hum, the ability to process words.

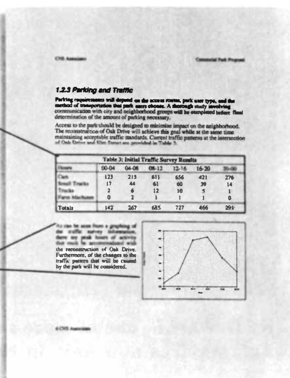
But, the mere processsing of words is hardly the challenge today. When you're creating a 20-page report, an 80-page proposal, a 200-page specification, or just a 2-page letter, you often have to work with graphics, equations, tables and more.

That's simply too much to ask of a product designed to process words. But, that's exactly what you can expect from Lotus Manuscript,™ a word processor that is truly a complete document creation system for technical professionals.

Very often documents require the mixing of text and graphics on the same page, graphics from Lotus® 1-2-3,® Symphony® and Freelance® Plus. With Manuscript you can import spreadsheets and charts, as well as diagrams and scanned images.

Spreadsheets, graphs and charts for 1-2-3® and Symphony® can easily be mixed with text on the same page

You can throw away your scissors and glue, cut and paste are a thing of the past.



1.2.3 Parking and Traffic

Parking requirements will depend on the access routes, park user type, and the method of transportation that park users choose. A thorough study involving communication with city and neighborhood groups will be completed before final determination of the amount of parking necessary.

Access to the park should be designed to minimize impact on the neighborhood. The reconstruction of Oak Drive will achieve this goal while at the same time maintaining acceptable traffic standards. Current traffic patterns at the intersection of Oak Drive and Elm Street are provided in Table 3.

Table 3: Initial Traffic Survey Results						
Hours	00-04	04-08	08-12	12-16	16-20	20-00
Cars	123	215	611	656	421	276
Small Trucks	17	44	61	60	39	14
Trucks	2	6	12	10	5	1
Farm Machines	0	2	1	1	1	0
Totals	142	267	685	727	466	291

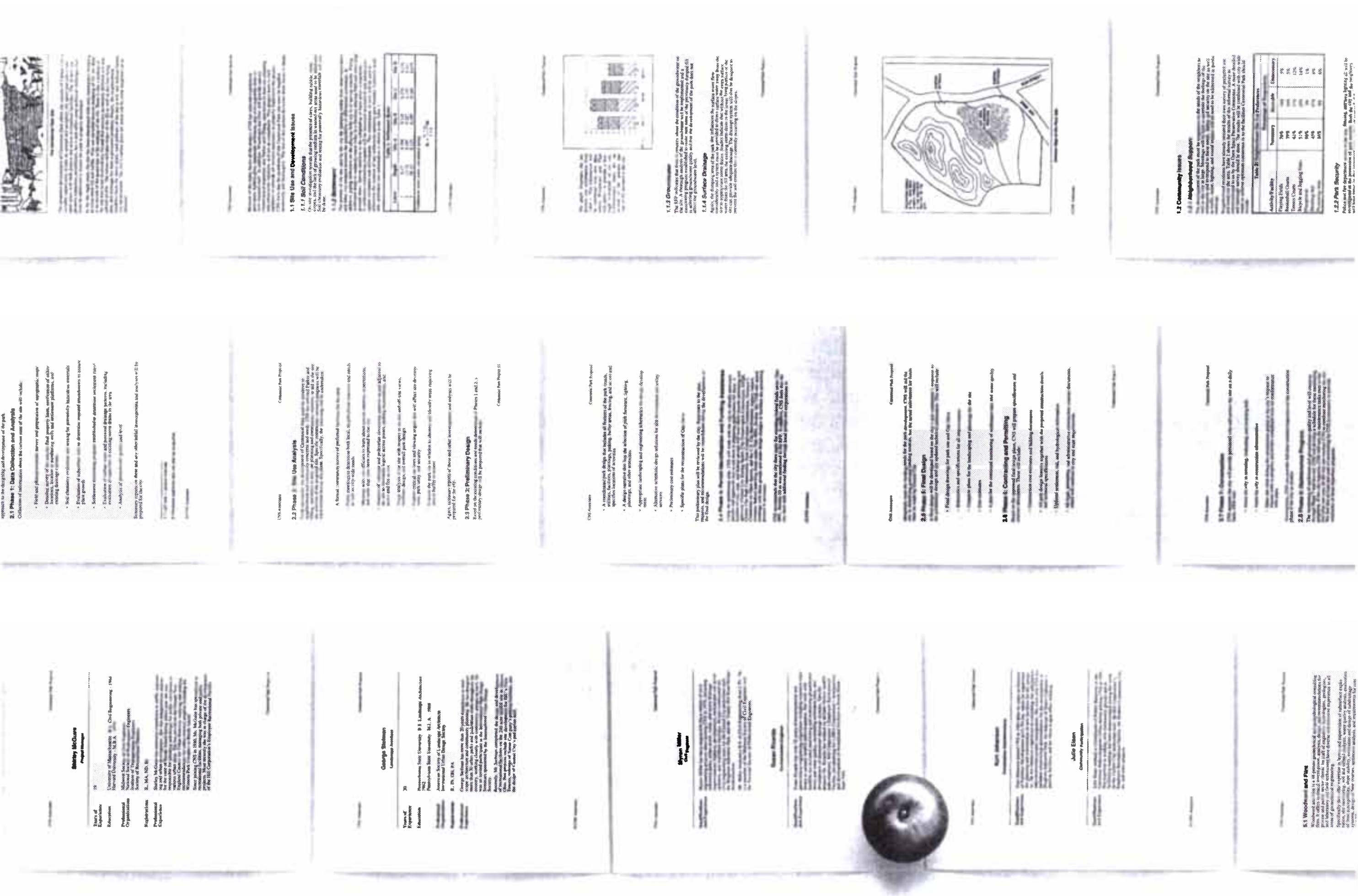
As can be seen from a graphing of the traffic survey information, there are peak hours of activity that must be accommodated with the reconstruction of Oak Drive. Furthermore, of the changes to the traffic pattern that will be caused by the park will be considered.

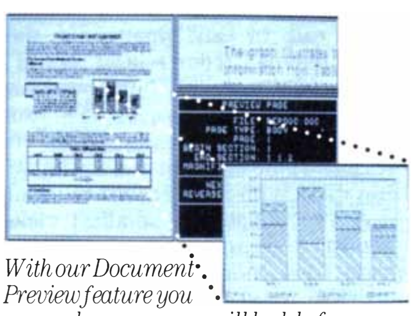
Another feature that goes beyond conventional word processing is Manuscript's integrated Outliner. When you collapse the document you're able to navigate the outline much easier than wading through the entire piece. You can move a page, or even an entire chapter with just a few keystrokes.

Because Manuscript understands the structure of your document, you can globally format an entire document or format by individual sections. Imagine changing all headlines from 14 point plain text to 16 point bold, by a simple menu selection.

Manuscript's Document Com-

ents instead of words.





With our Document Preview feature you can see how any page will look before it's printed. Zoom capability lets you take a closer look at graphics and equations.

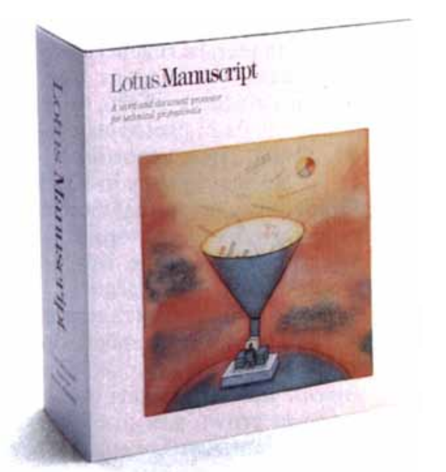
pare feature highlights changes between revisions for fool proof proof-reading.

Also, Manuscript automatically sizes and generates math equations as well as upper and lower case Greek symbols, diacritical marks and brackets.

With our powerful Print Formatter you have complete control over the look of your document, from position and size of graphics, to fonts and point sizes and more.

As for the quality of output, Manuscript takes full advantage of today's printing technology, from dot-matrix to laser, including PostScript® printers.

Manuscript is designed to work on most IBM® PCs and compatibles.* Its familiar 1-2-3 interface makes it easy to use. And our Manuscript evaluation kit makes it easy to try. For $10.00, you'll get a presentation disk, working software, and a tutorial manual. To get your evaluation kit, call 1-800-345-1043, ask for lot #BD-1450. Or, for more information, see your authorized Lotus Dealer, or write Lotus Development Corp., 55 Cambridge Parkway, Cambridge, MA 02142.



Lotus Manuscript™

mous. Last year the House of Representatives voted to reauthorize the act; in the Senate, however, a few members insisted that the act be amended to allow activities in their states, including a dam project in Texas that could wipe out the Concho water snake and strip-mining in Alabama that threatens the flattened musk turtle. When the senators failed to get the amendments, they blocked reauthorization. This spring, as red wolves take their first steps back into the wilderness, the 100th Congress is expected to hold hearings on reauthorization of the act.

Ingrained Genes

Genetic engineering has scored several recent successes in plants, such as petunias bearing an added gene that enables them to tolerate a weed killer, and tobacco plants endowed with genetically engineered immunity to a specific disease. Yet the most important food plants have resisted genetic manipulation, which in theory could increase their yield, nutritive value and hardiness.

A taxonomic boundary defines the problem: the plants for which successful methods of gene transfer have been devised are dicots, one subclass of flowering plants. Cereals, which in much of the world are the staple crops, all belong to the other subclass, the monocots. Work at a number of laboratories suggests that the taxonomic restriction can be bypassed: there are early signs that techniques for producing a genetically altered plant can also succeed in agriculturally important monocots.

In one of the current approaches protoplasts—plant cells from which the tough cell wall has been enzymatically digested away—are stimulated to take up foreign DNA in the test tube. A complete, genetically transformed plant can then be regenerated from the protoplast. For reasons that are not fully understood, efforts to regenerate monocots from protoplasts had generally failed. Now, according to an article in *Science,* groups in England and Japan have shown that rice plants can be grown from protoplasts—a result that should open the way to genetic manipulation in that grain.

The second approach to plant genetic engineering is founded on the peculiar talents of *Agrobacterium tumefaciens,* the soil bacterium that is the cause of crown gall disease. As it infects a plant *Agrobacterium* transfers a length of its own DNA (*T*-DNA) into the cells of its host. The *T*-DNA is incorporated into the plant's chromosomes and the genes it carries are expressed, inducing the formation of a crown gall. One can exploit this natural feat of genetic engineering by splicing a foreign gene into the *Agrobacterium's* *T*-DNA. The passenger gene is then incorporated into the plant's chromosomes, yielding transformed cells from which a complete plant can be regenerated. Only dicots are susceptible to crown gall disease, and so it was thought *Agrobacterium* could not transfer DNA to cereals.

By devising a sensitive test for *T*-DNA transfer, Nigel Grimsley, Thomas Hohn and Barbara Hohn of the Friedrich Miescher Institute in Basel and Jeffrey W. Davies of the John Innes Institute in Norwich, England, have shown otherwise. Into the *T*-DNA of *Agrobacterium* the workers inserted multiple copies of the DNA from a plant pathogen, maize streak virus (MSV). The group inoculated maize plants with the altered bacterium. Within two weeks, the investigators report in *Nature,* yellow streaks on the leaves testified to widespread MSV infection. Because naked DNA from the virus cannot infect maize cells on its own, *T*-DNA transfer must have taken place, carrying the viral DNA into the maize cells.

The demonstration that *Agrobacterium* can transfer DNA to maize is only a first step toward making the bacterium a vehicle for genetic engineering in monocots. The investigators do not know, for example, whether the transferred DNA was integrated (as it is in a dicot) into the maize chromosomes—a necessary step if a plant is to be given a new heritable trait.

Workers at the Max Planck Institute for Plant Breeding Research in Cologne have accomplished just such integration in rye, another monocot, by devising an entirely new approach to plant genetic engineering. They describe their strategy in *Nature.* In collaboration with other colleagues, one of the workers, Alicia de la Peña, had found that two weeks before the precursor cells of pollen begin their final series of cell divisions in male rye plants they become singularly permeable to certain foreign molecules. Hoping the cells might take up even a molecule as large as DNA, de la Peña, Horst Lörz and Josef Schell tested a plasmid (a circular DNA molecule) that included a gene conferring resistance to kanamycin, an antibiotic that is also lethal to plant cells. They injected the plasmid into the young floral shoots of rye plants.

When the floral shoots matured, the investigators cross-pollinated the injected plants. If the injected DNA had transformed any of the germ cells, the group reasoned, at least a few seeds resulting from the cross-pollination might sprout into kanamycin-resistant plants. To test for kanamycin resistance de la Peña and her colleagues germinated more than 3,000 seeds on a medium containing the antibiotic. Seven seedlings survived. Tests for the enzyme encoded by the kanamycin-resistance gene and for the foreign DNA itself indicated that in two cases the resistance did indeed signal a genetically transformed plant.

The Max Planck workers point out that germ cells in other cereals develop much as they do in rye. They predict that their simple technique of directly injecting foreign DNA into plants at the appropriate stage of floral development will also succeed in other grains. With their work and that of other groups the promise of plant genetic engineering comes a step closer to reality.

Bringing Up Baby

From little eggs mighty dinosaurs grew—or did they? Brontosaurs, which at more than 25 tons were among the largest of all dinosaurs, did not hatch from football-size eggs as some other dinosaurs did; they entered the world viviparously, as big as a full-grown pig, according to Robert T. Bakker of the University of Colorado at Boulder. Bakker, who has proposed that dinosaurs were warm-blooded, thinks brontosaur mothers may have nurtured their single strapping newborns instead of leaving them to fend for themselves, as reptiles do their hatchlings.

In support of his hypothesis, Bakker points out that the largest dinosaur eggs known are only about eight inches in diameter; beyond that size the shells would have been too thick to break or to allow oxygen in the air to pass through to the embryo. The mortality rate of tiny, unattended dinosaurs hatched from a large group of eggs would probably have been high, and indeed paleontologists have found mass graves of hatchlings from such species as the duckbill and the finback. But there is no fossil record of brontosaur hatchlings; the smallest clearly identified infant brontosaur skeletons were from animals that would have weighed about 300 pounds.

Female brontosaurs had birth canals that could easily accommodate a fetus this size; Bakker notes that the pelvic cavity of fossilized females was nearly wide enough "to drive a Volkswagen bug through." He believes a relatively long gestation period followed by the birth of a single baby rather than a large brood would have led to "high parental investment." For this reason, he suggests, brontosaur moth-

M I T S U B I S H I

Soothe your nervous system.

There are those who measure distance in miles. While others choose to measure distance in terms of time.

At Mitsubishi Motors, we take an altogether different view. We believe that the best measure of distance in a luxury sedan is the distance you never notice.

And to prove it, we created the Mitsubishi Galant.

Galant's engineered to deliver a unique driving experience. Patiently crafted to be an island of calm amid the storm of demands success often brings. Galant's onboard computer, ETACS IV™, monitors all secondary functions from power windows and door locks to a new theft-deterrent system. Its electronically fuel-injected 2.4-liter engine responds with a silent surge of power. There's even an available ECS™ electronically controlled suspension that self-adjusts to changing road conditions.

Of course a sedan this sophisticated demands luxury as well as performance, so inside you'll find a haven of comforts and appointments. Adjustable front bucket seats, in velour or optional leather. Electronic power steering for smooth, confident handling. An available six-speaker AM/FM cassette stereo system with graphic equalizer and remote controls mounted on the steering wheel. And to ensure that passengers in back are every bit as comfortable, Galant offers a unique reclining rear seat with adjustable headrests.

The Mitsubishi Galant. Road car performance in a luxury sedan. Patiently engineered to soothe your nervous system and to keep an impatient world at a comfortable distance.

Call 1-800-447-4700 for your nearest dealer.

Mitsubishi. Perfecting the Experience.™

MITSUBISHI MOTORS

G A L A N T

ers would have been likely to tend their young rather than abandoning them. Brontosaur "trackways" found in Texas and Colorado, which show mingled small and large footprints, and the lack of isolated footprints of juvenile brontosaurs, offer further evidence of a close relation between offspring and parents, Bakker says.

Other scholars are not entirely convinced. Walter P. Coombs of Western New England College observes that the fact that juvenile and adult brontosaurs traveled together does not mean mothers and infants had a close relationship. Coombs believes Bakker's live-birth thesis needs "smoking-gun evidence": a fossil of a female brontosaur with a large fetus inside would do.

Bakker replies that such evidence may exist. In 1901 the remains of a dinosaur that would have weighed about 150 pounds were found scattered among the bones of an adult brontosaur. At the time, paleontologists categorized the small fossil as a dinosaur of another species, an elosaurus, but Bakker argues that it may have been an embryo carried inside the adult. Bakker, whose theories are propounded in a new book, *The Dinosaur Heresies,* agrees with Coombs that his case for live birth is a tentative one. Yet he is confident that evidence for it will turn up, since brontosaur fossils are found every year.

Might other large dinosaur species, the *Tyrannosaurus rex,* for example, also have given live birth? Probably not, Bakker says. Unlike the brontosaur, *T. rex* had a tiny birth canal; moreover, the species was apparently closely related to birds. "I suspect," Bakker muses, "that the guy had a bird style of reproduction."

Sopranos of the Skies

The effect of a whiff of helium on the pitch of the human voice is familiar: basso profundos become reedy tenors. Songbirds breathing helium, on the other hand, seem to a naive listener to be only slightly affected; the most audible difference in their chirps is not a rise in pitch but a drop in volume. This observation, which was made in experiments more than 20 years ago, has contributed to a belief among ornithologists that the physiological mechanisms shaping birdsong are far simpler than the mechanisms from which human phonation springs. Investigators thought birds merely vibrate their syrinx, the vocal organ, and do not modify this basic "source sound" by altering the shape of the vocal tract, as human beings do.

Stephen Nowicki of Rockefeller University challenges that view. With acoustic analyzers more sensitive than those available to earlier investigators, he discovered that when chickadees, sparrows and other birds sing in helium, high-frequency overtones emerge that are difficult for the casual listener—and for less sensitive instruments—to detect. Nowicki's discovery, reported in *Nature,* suggests that, contrary to what has been thought, songbirds are in fact much like miniature opera singers, working hard to shape every utterance.

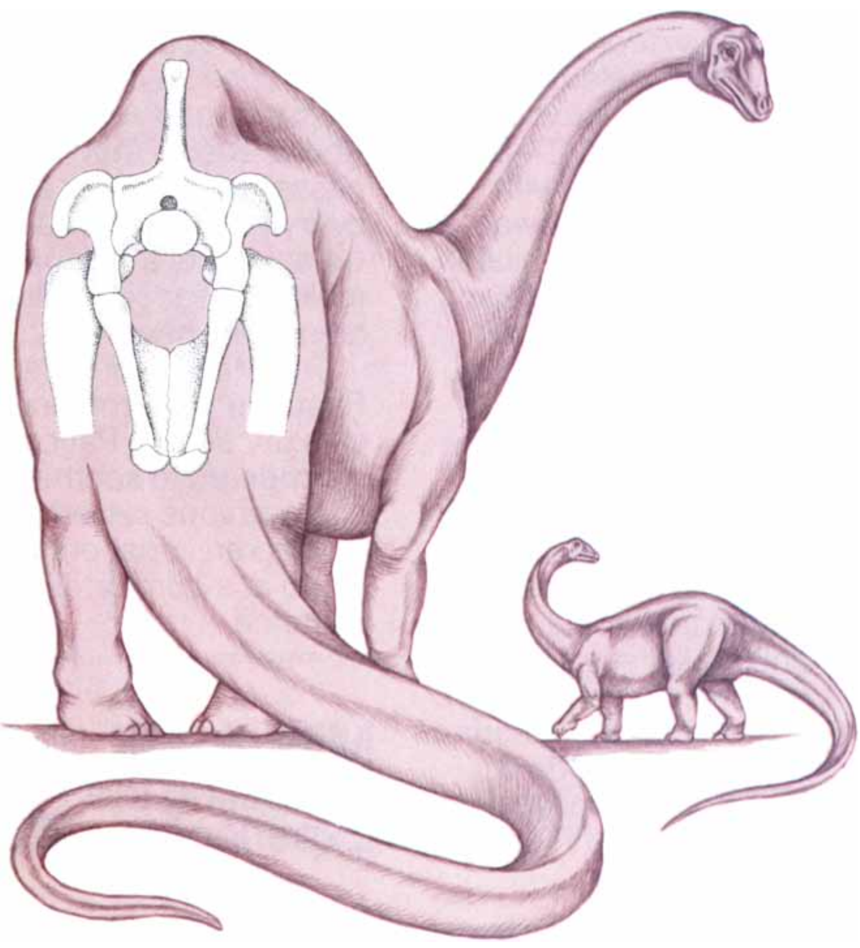
When a soprano sings a high C, her vocal cords actually produce a broad band of frequencies. The soprano, however, manipulates her vocal tract—which includes the trachea, larynx, throat and mouth—so that it resonates at the same frequency as high C and thereby amplifies it; at the same time the vocal tract acts as a filter, absorbing other tones.

If the soprano inhales helium, her voice seems to rise in pitch not because her vocal cords vibrate faster in the less dense atmosphere (they do, but only slightly); rather, because sound travels almost twice as fast through helium as it does through nitrogen, the acoustic properties of the vocal tract change so that it resonates with and amplifies higher-frequency tones.

Nowicki's discovery of the high-frequency overtones suggests that, like the soprano's high C, each apparently pure utterance of a songbird's syrinx actually includes several tones; birds emphasize one tone in preference to others by extending or shortening the trachea, by constricting the larynx or by flaring the throat or beak.

There are significant differences, Nowicki notes, between the vocal physiology of songbirds and that of humans. Most notably, the syrinx, unlike human vocal cords, has two parts that can simultaneously produce two distinct sounds. Nevertheless, Nowicki says, his helium experiments show that because the sounds of both songbirds and human beings arise from a "complex, coordinated ballet of muscle activity," they are more alike than had been thought.

Nowicki points out that he undertook his experiments at the urging of Peter Marler, who heads bird studies at Rockefeller's Millbrook field station and is a pioneer in the field. It was a graduate student of Marler's who, in 1966, first recorded the songs of birds



COMMODIOUS PELVIC CAVITY of the female brontosaur suggests that she may have given birth to single fetuses weighing perhaps 300 pounds rather than laying small eggs.

in helium and was unable to discern the overtones hidden in their tunes.

MEDICINE

"A Pothole System"

In a major assessment of the state of national health the Reagan Administration has declared that "substantial progress" has been made toward 226 national goals established at the end of the Carter Administration. Like its predecessor, the Reagan Administration has committed itself to achieving the goals by 1990. The implication that all—or at least much—is well on the health-care scene was disputed on Capitol Hill. There a Senate committee heard that health care in the U.S. is "a pothole system because it damages our people and shatters their lives."

The Administration report, issued by the Department of Health and Human Services, is titled *The 1990 Health Objectives for the Nation: A Midcourse Review.* The report foresees success in achieving 110 of the 226 goals. They include better control of high blood pressure, less morbidity from cardiovascular disease, the prevention of unintentional injuries, a reduction in the proportion of adult smokers, lower alcohol consumption and less cirrhosis, and a reduced death rate from automobile accidents. A major goal that seems unlikely to be reached is an infant mortality rate of no more than nine deaths per 1,000 live births. Other goals that are apparently out of reach are a lowered rate of teen-age pregnancy, a reduction in the rate of sexually transmitted disease and notably increased public participation in nutrition, fitness and exercise.

It would be difficult to attribute such gains as have been made to participation in the health-care system; some are clearly the result of "buckling up" and other preventive measures. Furthermore, there are strong assertions that the health-care system is deteriorating. In hearings held by the Senate Labor and Human Resources Committee, Joseph A. Califano, Jr., who was Secretary of Health, Education, and Welfare in the Carter Administration, cited the "unprecedented" fact that "fewer people than ever are covered by [health] insurance." The explanation, he said, is the rise in service-industry jobs that provide little or no health insurance. This is one reason for the "pothole system" in health care, he said.

Robert M. Ball of the Center for the Study of Social Policy cited another reason, telling the committee that Medicaid and Medicare are inadequate. Medicaid, he said, "is available to less than 50 percent of the population living below the rock-bottom level of officially defined poverty," which in 1985 was $10,989 per year for a family of four. Medicare, he added, "now pays only a little more than 40 percent of the total health-care costs of the elderly," who "are paying as much today as they were before Medicare first came in."

The people who fall through such cracks in the health-care system—the poor, the elderly and the uninsured or underinsured—wind up in public hospitals, according to some students of the system. One such hospital, the Santa Clara Valley Medical Center, has seen admissions rise 61 percent since 1983. Robert Sillen, the executive director of the hospital, considers the phenomenon symptomatic of a nationwide problem. "We are on the borderline of a health-care crisis," he said. "In the next two years or so, people will start asking how we, the richest society in the history of the world, can treat the most vulnerable segments of our population in a totally unacceptable manner."

Arnold S. Relman, editor of *The New England Journal of Medicine,* has similar views. "It is a paradox of health care in the U.S.," he said, "that even as it is becoming ever more sophisticated, it is also becoming ever more inequitable."

Breast Cancer: neu *Clue*

Breast cancer patients whose tumor cells contain extra copies of a particular oncogene, or cancer gene, appear to have a dimmer prognosis than other patients, according to a group of investigators from the University of California at Los Angeles School of Medicine and the University of Texas Health Science Center at San Antonio. The news is not all bad. The discovery may provide a new way to identify patients who need the most intensive treatment, and it may also help to elucidate how the disease develops.

Dennis J. Slamon of the U.C.L.A. School of Medicine and his colleagues became interested in the oncogene, HER-2/*neu,* after other investigators reported that it is amplified in about 10 percent of the cell lines derived from human breast cancers. Work by another group at U.C.L.A. alerted Slamon to the potential role of gene amplification as a guide to prognosis; those colleagues found a correlation between extra copies of the oncogene N-*myc* and reduced survival time in children suffering from neuroblastoma, a rare nerve-cell cancer.

By examining tissue from 103 primary breast cancers, Slamon's group first found that patients whose disease had spread to three or more axillary, or armpit, lymph nodes were the most likely to have amplified HER-2/*neu* genes in their tumor cells. Because disease in the nodes is currently the most potent clinical indicator of relapse and survival times, the workers concluded that HER-2/*neu* might well serve as a a strong predictor of future health.

The investigators found a direct correlation between degree of amplification and prognosis in a separate group of 86 patients whose cancer involved the lymph nodes: the greater the number of HER-2/*neu* copies, the shorter the time to relapse and the shorter the survival time, particularly for individuals whose cells had more than five copies of the gene.

Slamon and his colleagues are now studying the link between HER-2/*neu* amplification and prognosis in breast cancer patients whose lymph nodes are free of disease. They reason that individuals who have cancer in the lymph nodes (women with relatively more advanced disease) undergo radiation therapy or chemotherapy routinely after the initial tumor is removed; the new prognostic information provided by gene amplification might not significantly alter the treatment of such patients. In contrast, individuals with "clean" lymph nodes are generally assumed to have a good chance of avoiding a cancer recurrence and are spared such treatment. Unfortunately, Slamon notes, cancer does recur in perhaps 25 to 30 percent of the node-negative individuals. If multiple copies of HER-2/*neu* predict that a relapse is likely or that survival time may be shortened, physicians will finally have a way to pinpoint even the node-negative patients who need aggressive treatment. Results are expected later this year.

If overproduction of a known oncogene correlates with reduced life expectancy in cancer patients, the gene is probably contributing to the cancer. How might HER-2/*neu,* which is known to cause nerve-cell cancer in rats, lead to breast cancer in humans? The oncogene appears to encode a cell-surface receptor for an as yet unidentified growth factor. By binding specific growth factors, such receptors direct the cell to multiply. The investigators speculate that amplification of the HER-2/*neu* gene could well result in the production of extra receptors. These might bind excess growth factor, causing the affected cells to grow abnormally fast. Other explanations are also possible. In any case, if excessive binding and receptor activity is a culprit in breast cancer, identification


Mon.
Tues.
Thurs.
Fri.
The week David Ansen went to the movies and wrote a story on old age, high art and the bottom line.
When Newsweek sends its critic, David Ansen, to a movie —he goes to a movie.
He enters its beckoning world with the intensity of a foreign correspondent covering a war. And he prepares his dispatches with an all-seeing eye that can spot nuance in a disaster and romance in a stone.
Ansen's reviews are welcomed by Newsweek readers because they're so much more than thumbs up or thumbs down. He climbs inside the movie to tell you why it was made. And how it was made. And whether the world would have missed it if it hadn't been made.
When he writes a story on Clint Eastwood he helps you see him for what he really is. Not just a lucky survivor of spaghetti westerns but as a man who has merged the characters he portrays into the character he is: "An American icon."
This isn't Hollywood hoopla. This is sociology. Anthropology. Economics. Always infor-


© NEWSWEEK 1986
Wed.
A scene from the motion picture, "RAN." Directed by Kurosawa, inspired by "King Lear."
Sygma
Sat.
mative and always entertaining.
When he pays tribute to Akira Kurosawa, the 75-year-old master of Japanese film, Ansen treats him not only as a genius of the Samurai movie but as a sharp interpreter of contemporary life. A wise teacher who loves the world enough to be saddened by it.
Ansen sees Kurosawa as a national treasure rejected by his own country—an aged wanderer seeking support.
David Ansen belongs in the movies. And because of the special way he sees them—the bridging of their cultural, business, social and international significance—he belongs in Newsweek.
That's where you'll find him.
Newsweek
Why it happened. What it means.

of the putative binding factor could lead to the development of a substance that blocks its activity and, possibly, to a new treatment for a major killer.

TECHNOLOGY

Beeping the Faith

It is said that in Islam science is the handmaiden of religion. Indeed, many early astronomical and mathematical techniques were devised specifically in order to ensure the proper execution (in place and time) of Moslem observances. Two articles in the *Journal of the Royal Astronomical Society of Canada* show that science and technology continue to serve the changing needs of the modern, globe-trotting observant Moslem.

One of a traveling Moslem's first needs on reaching a strange latitude and longitude is to find out when to say the five daily prayers (the times of prayer are based on astronomical phenomena whose hours of occurrence vary from one locale to another) and what direction to face while saying them. Now the information is in easy reach, if he has brought along his Prayer Times Clock.

The clock, a digital affair with a liquid-crystal display and powerful microelectronic insides, is described by David M. Stokes of the Royal Astronomical Society's Kingston Center. On its case, in addition to the conventional "set" and "alarm" buttons, are buttons marked (in Arabic and English) *fajr* (dawn), *shurūq* (sunrise), *ẓuhr* (noon), *'aṣr* (afternoon), *maḡrib* (sunset) and *'ishā* (night). (*Fajr* and *'ishā* represent astronomical twilight, when the sun is 15 degrees below the horizon.) Although there are five prayers, the clock has six buttons because the first prayer is said between dawn and sunrise. After a traveler has programmed the clock with his latitude and longitude, pressing a button gives the time of the astronomical event indicated by that button. The clock also sounds an alarm five minutes before each event and at the time of the event, and it displays the name of the prayer to be said.

But which way to face? The clock solves that problem also. Pressing any button twice displays the compass bearing of Mecca. The price of the clock, says Stokes, is "$30.00 to $45.00 (Canadian) depending on your ability to bargain in Arabic."

The traveler's next problem is to find out the day on which each new month begins, which is particularly important in determining the dates of holidays. The difficulty is that the month begins on the evening a new moon is first visible. This is an inherently local phenomenon, and so it has been hard to devise an international lunar calendar that could be referred to by travelers in foreign lands. The result is confusion.

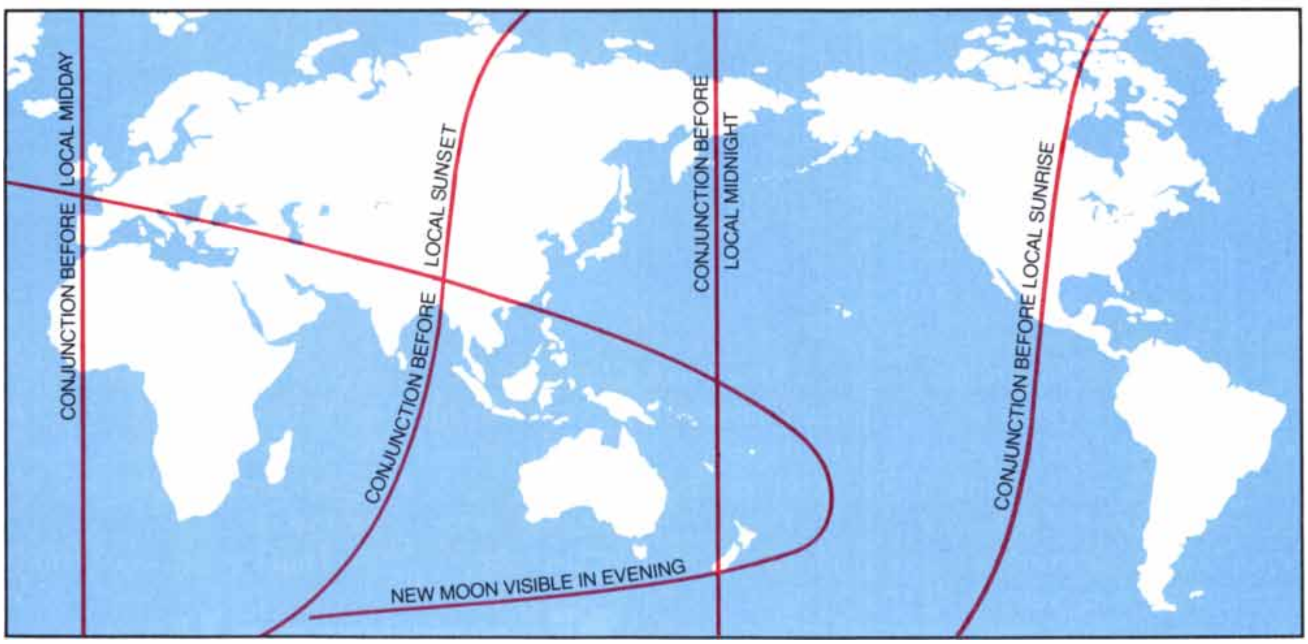
Mohammad Ilyas of the University of Science of Malaysia reports progress toward a solution. The answer, according to Ilyas, is to find an international lunar date line: a line marking the easternmost point, at each latitude, from which the new moon is first visible in the evening. Anywhere to the west of the date line, the new moon will become visible by evening and the month can be said to have begun. To the east of the date line the month will not begin until the next solar day. As Ilyas points out, the position of the date line changes each month.

Calculating the position and shape of the date line is not a simple matter. Ilyas first had to devise a way of finding the probability that the new moon would be visible at a given point on the earth's surface. He also determined the shapes date lines would have in the lunar calendars followed by other religions. In some cases, for example, the month is said to begin if conjunction (the alignment of the sun, moon and earth that occurs at new moon) takes place before local noon or before local sunrise.

Ilyas' system of date lines makes it possible to produce a fully global calendar that ensures faithfulness to local astronomical phenomena. He thus solves the quandary of isolated communities or travelers who do not have a "basic understanding of religio-scientific requirements."

Light Switch

An inquiry into how the semiconductor gallium arsenide responds to light has uncovered a phenomenon that may one day find application in an optical computer. Such a machine, in which information would be transmitted by beams of light rather than by



DATE LINES for an international lunar calendar mark, at every latitude, the easternmost point at which a new lunar month begins on a certain solar day. The shape and position of a line depend on the local astronomical phenomenon that determines the start of the lunar month. The position of the date line changes every month. The lines here refer to the new moon of January 10, 1986.

electric currents, would be faster than its electronic counterpart; it might also be better suited to parallel information-processing strategies that would further augment its capacity. To build an optical computer, however, one must first have the optical equivalent of a transistor: a device in which one light signal would switch another on or off. Research groups around the world are pursuing various approaches to this problem, and prototypes have already been designed.

A novel approach, albeit one that is still far from application, is suggested by the work of André Mysyrowicz and his colleagues at the École Polytechnique in Paris. They report in *Physical Review Letters* that an ultrafast-pulsating laser can produce ultrafast shifts in the wavelength at which gallium arsenide absorbs light from another laser. Such shifts could serve as the equivalent of the "on" and "off" states of a transistor.

A semiconducting material absorbs light when photons cause electrons in it to jump to a higher energy level. In making the jump the negatively charged electrons leave behind them positively charged "holes" in the crystal lattice. The holes, which can be treated as particles, exert an electrostatic attraction on the electrons. Indeed, the two tend to form bound states, called excitons, that are analogous to ordinary hydrogen atoms. Each exciton represents the absorption of one photon.

To create excitons the laser photons must have enough energy to lift the electrons in the semiconductor across the "band gap" that separates one quantized energy level from the next. At the light wavelength that corresponds precisely to the band-gap energy plus the binding energy of an exciton (the resonance wavelength), the formation of excitons and hence the absorption of light by the semiconductor reach a sharp peak. Wavelengths immediately to the sides of the peak pass through the semiconductor and are detected.

In the experiments reported by the French workers, two lasers, both pulsating at time intervals of less than a trillionth of a second, were shone onto a composite semiconductor consisting of alternating thin layers of gallium arsenide and gallium aluminum arsenide. The "probe" laser emitted a weak beam across a broad spectrum that included the resonance wavelength of the semiconductor: about 800 nanometers, in the near-infrared range. The "pump" laser emitted an intense beam at a slightly longer wavelength at which the photons do not have enough energy to create excitons.

Mysyrowicz and his colleagues report that when the probe laser alone was on, the absorption spectrum of the semiconductor showed the expected peak at the resonance wavelength. During each pulse of the pump laser, however, they observed a distinct shift of the absorption peak to a shorter wavelength and a lower amplitude. At high pump intensities the peak was nearly flattened.

To account for their results the investigators offer the following hypothesis. Photons from the pump laser, although unable to create excitons themselves, can couple with excitons generated by the probe laser. The energy necessary to create such a "dressed" exciton is greater than that necessary to make an ordinary exciton. As a result the absorption of probe-laser light by the semiconductor shifts toward the shorter-wavelength, higher-energy end of the probe band. Because fewer excitons are being generated, the absorption peak is also flattened.

Since the change in the absorption peak is caused by the pump photons, it does not last much longer than the pump-laser pulse: about a trillionth of a second. After that the absorption of probe light by the semiconductor returns to normal. Therein lies the technological promise of the phenomenon. In principal a pump laser, by shifting the absorption characteristics of a semiconductor, could serve as the equivalent of the base on a transistor, switching the signal emitted by the probe laser on and off every trillionth of a second (many times faster than the switching speed of the fastest transistors). A number of obstacles, however, still stand in the way of a practical exploitation of the effect, including the enormous intensity required of the pump laser.

Clearing the Air

Nitrogen oxides produced by the combustion of hydrocarbon fuels are among the most troublesome of atmospheric pollutants. Nitric oxide and nitrogen dioxide (NO and NO_2) are converted into ozone, which is largely responsible for the smog blotting the skyline of such cities as Los Angeles and Denver, and into nitric acid, which contributes to acid rain.

Catalytic converters have reduced emissions of nitrogen oxides from gasoline-fueled automobile engines, but until an investigator at the U.S. Department of Energy's Sandia National Laboratories in Livermore, Calif., addressed the problem no effective means seemed at hand for eliminating the oxides from the exhaust of two other major polluters: diesel engines and coal-burning plants. The investigator, Robert A. Perry, has developed a process, called rapid removal of nitrogen oxides (RAPRENOX), that may be able to virtually eliminate the compounds from diesel and coal exhaust.

As Perry and a co-worker, Dennis L. Siebers, report in *Nature,* RAPRENOX entails routing exhaust past the common nonpoisonous compound cyanuric acid, which is ordinarily used as a stabilizer of chlorine in swimming pools. When cyanuric acid crystals are heated to a temperature of 350 degrees Celsius or more, they release isocyanic acid gas. The gas reacts with nitrogen oxides in the exhaust, leaving nitrogen, carbon dioxide, carbon monoxide and water. According to Perry, one pound of cyanuric acid, at a cost of $1, can eliminate an equivalent amount of nitrogen oxides. In the only test of the process so far, Perry and Siebers successfully removed 99 percent of the nitrogen oxides in exhaust produced by a one-cylinder diesel engine.

RAPRENOX has some significant advantages over other techniques for removing nitrogen oxides from emissions, Perry maintains. Diesel-exhaust scrubbers, as they reduce the oxides, cause levels of carbon particulates to increase; RAPRENOX, he says, does not. Methods for reducing nitrogen oxides from coal plants generally require extensive modification of the combustion system. Perry thinks the exhaust system of plants could be retrofitted with a RAPRENOX-based system more easily and cheaply.

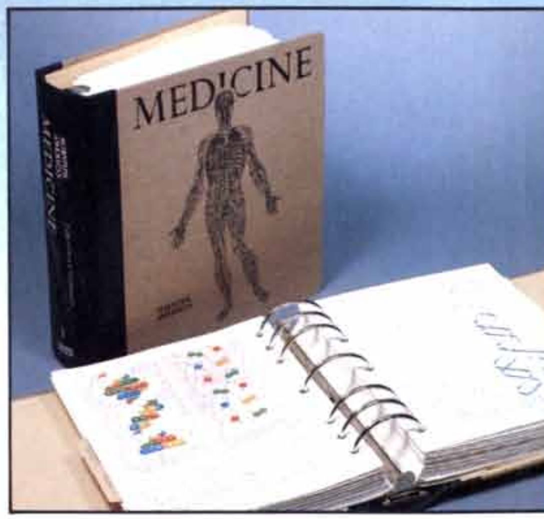
The Government, while reserving the option to exploit the process for its own purposes, has transferred the patent rights to Perry, who plans to leave Sandia to develop RAPRENOX for commercial applications. He has found considerable interest in his invention, notably in smog-afflicted Los Angeles. After a briefing from Perry, California's South Coast Air Quality Management District Board recommended that state agencies help to speed the development of RAPRENOX. One board member, Thomas Heinsheimer, estimates that outfitting diesel engines in the Los Angeles basin with RAPRENOX-based devices could eliminate almost 20 percent of the 900 tons of nitrogen oxides spewed daily into the atmosphere. (The rest comes from cars, industrial processes and other sources.)

Another pollution expert is more cautious about RAPRENOX' potential for coal-burning plants. "It's one thing to make it work in a one-inch-diameter pipe," John S. Maulbetsch of the Electric Power Research Institute in Palo Alto, Calif., says. "It's another thing to make it work in a 20-foot duct filled with fly ash."

OUR SYSTEM: a rewarding, efficient way to keep yourself up-to-date—and save hundreds of hours of your time for patient care.

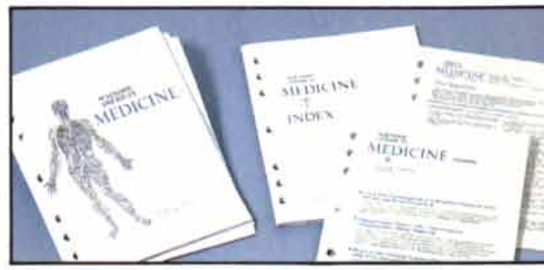
A comprehensive, 2,300-page text in two loose-leaf volumes, incorporating the latest advances in medical practice as of the month you subscribe.

This superbly designed, heavily illustrated resource, called "the best written of all [the internal medicine] books" by *JAMA* (251:807, 1984), provides a practical, comprehensive description of patient care in 15 subspecialties. And, because the text is updated each month, the clinical recommendations reflect all the current findings. A practice-oriented index and bibliography of recent articles further enhance the efficiency of the text.



Each month, six to nine replacement chapters to update your text *plus* new references, a four- to five-page news bulletin, and a completely new index.

You'd have to read hundreds of journal pages each month—and memorize the contents—to get the same information SCIENTIFIC AMERICAN *Medicine* contains. With our updated text, you read only the information you really need. Our authors, largely from Harvard and Stanford, sort through the literature and monitor developments, incorporating the significant advances into our chapters.



At no additional cost, a 32-credit CME program, to save you valuable patient-care time and the expense of attending review courses.

Earn 32 Category 1 or prescribed credits per year with our convenient self-study patient management problems; each simulates a real-life clinical situation. Choose either the complimentary printed version or, at a modest extra charge, the floppy-disk version for your IBM® PC/PC Jr., Macintosh™, or Apple® IIe or II+ (or compatible).

Your subscription includes:
the two-volume, 2,300-page loose-leaf text
and, each month, six to nine replacement chapters,
a newsletter, new references,
and a completely revised index.

Or call toll-free: 1-800-345-8112
(in Penn. 1-800-662-2444)

CINE

SCIENTIFIC AMERICAN MEDICINE
415 MADISON AVENUE, NEW YORK, N.Y. 10017

☐ **Yes,** I'd like to try the SCIENTIFIC AMERICAN *Medicine* system. Please enter my subscription at a first-year price of US$245*.

☐ Also enroll me in the CME program. I'd prefer:
- ☐ the printed version at **no additional charge.**
- ☐ the floppy-disk version at an additional $US97.*
  Computer type or compatible:
  - ☐ IBM® PC/PC Jr. (128K RAM, DOS 2.0/2.1)
  - ☐ Macintosh™ (512K RAM)
  - ☐ Apple® IIe (64K RAM)
  Apple® II + (64K RAM) 80-col. card by: ☐ Apple® ☐ Videx®

☐ Enroll me in the floppy-disk CME **only** at US$152.* (Note model above.)

☐ Check enclosed* ☐ Bill me

Signature ______

☐ VISA ☐ MasterCard Exp. Date ______

Account No. ______

Name ______

Specialty ______

Address ______

City ______ State ______ Zip Code ______

*Add sales tax for Calif., Ill., Mich. or N.Y. Allow 8 weeks for delivery. Add US$10 for shipping to Canada. IBM is a registered trademark of International Business Machines Corporation. Apple is a registered trademark of Apple Computer, Inc. Videx is a registered trademark of Videx, Inc. Macintosh™ is a trademark of McIntosh Laboratory, Inc. and is used by Apple Computer, Inc. with its express permission.

Monoculture

This practice, which entails growing the same crop on the same land repeatedly, has certain advantages for farmers but may not always be good agronomy

by J. F. Power and R. F. Follett

In 1907 a typical farm in the U.S. Corn Belt consisted of about 65 hectares (160 acres), on which the farmer grew corn, oats and hay, usually taking care to rotate the crops among his fields from year to year. He also kept horses, cattle, pigs and poultry, which contributed to the operation in various ways. The family grew its own vegetables and kept an orchard. Virtually all the labor was supplied by the farmer and his family, sometimes with the help of a hired man. The farm's sales consisted mainly of meat, eggs and milk. The family's purchases consisted mainly of wagons, harness, a few simple pieces of farm machinery, cloth and thread, and such items of food as flour, salt, sugar and spices.

Today a typical Corn Belt farm has from 200 to 300 hectares, on which the farmer grows mainly corn, rotated every few years with soybeans. Livestock are either absent or few in number. Although most of the labor is still done by the family with occasional hired help, many specialized services are purchased. They include equipment repair, soil testing, fertilizer application and crop spraying. Sales consist almost exclusively of the crops grown. The family's purchases have expanded to include large farm equipment, farm chemicals, almost all the seed for the crops and essentially all the needs of the household. Indeed, the owner may not be a family now; it is likely to be a large corporation.

When corn follows corn, this farm is practicing monoculture: growing the same crop on the same land through at least two crop cycles. The practice has become widespread in the U.S. over the past 80 years and is one of the major contributors to the ability of the American farm system to serve a market that extends far beyond the nation's borders. An example of monoculture is continuous corn, meaning corn grown on the same land year after year. About 21 percent of the corn grown in the Corn Belt follows a crop of corn. In many areas wheat is grown one year and the field is kept fallow the next year; that pattern too can be viewed as monoculture. It is estimated that about 50 percent of the wheat produced in the U.S. is grown as a monoculture; the same is true for sorghum. About a third, or $10 billion, of the income U.S. farmers receive from corn, wheat and sorghum production is the result of monoculture. A similar amount of income is derived from limited rotations such as corn-soybeans.

The chief forces driving the shift toward monoculture were mechanization, the improvement of crop varieties and the development of chemicals to fertilize crops and control weeds and pests. Monoculture offers a number of advantages to the farmer, but one can question whether it is altogether good for the land and the agricultural economy.

Reasons for the Trend

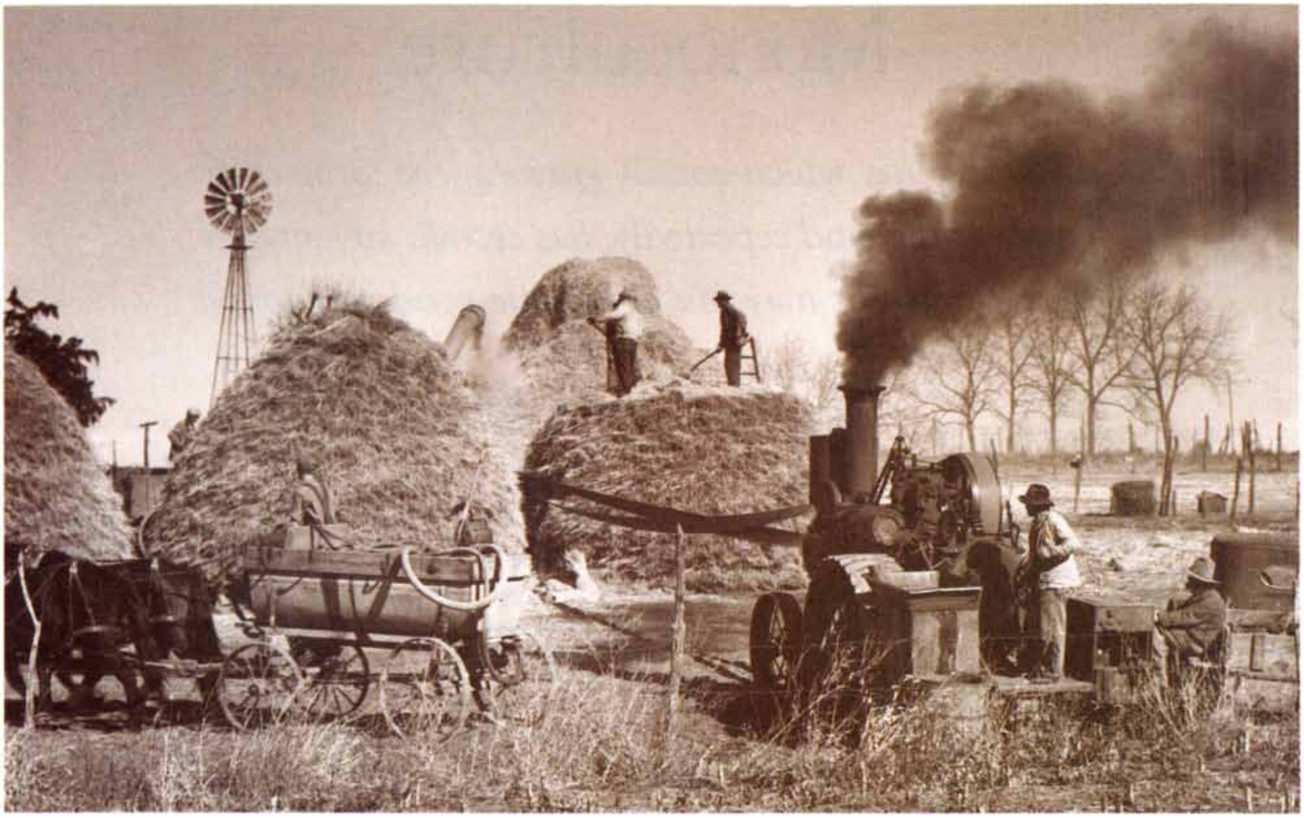
Mechanization, the first of the three forces pushing farmers toward monoculture and limited rotation, began in the 19th century with the development of the steam engine and the steel plow. By 1920 they had served to plow up much of the native prairie in the Great Plains. The steam engine had its limitations, however, because it was extremely slow, heavy and limited in maneuverability. Gasoline tractors began to appear after 1900. By the end of World War I they were readily available; by the end of World War II most of the commercial farms in North America had switched to them.

This conversion had several important effects on agriculture and cropping systems. No longer was it necessary to devote from 20 to 30 percent of a farm's cultivated land to the production of feed for draft animals. The land shifted out of feed crops was converted to the production of grain, often yielding more than could be consumed by the livestock remaining on the farm. As a result the farm derived an increasing percentage of its income from the sale of grain. With more money coming in, the farmer was able to expand his off-farm purchases to include larger and costlier farm machinery, together with the fuel to run the equipment and repair services to maintain it. Finally, because one operator could cover more land in a day with a tractor than with horses, farms became larger.

The size of tractors and other farm machinery has increased notably over the past two decades. In addition more specialized agricultural equipment is available: combines for corn, planting machines for fields where little or no tilling is done, high-clearance tillage equipment, center-pivot irrigation rigs and many others. These changes have further increased the size of farms because one person can farm more land than ever before. With a heavy investment in specialized machinery the farmer has a strong incentive to grow only the crop for which the machinery is designed.

The improvement of crop varieties (the second force pushing farmers toward monoculture and limited rotation) has proceeded along two lines: genetic advances in existing crops and the introduction of new crops. The first efforts, which began more than 100 years ago, consisted mainly of choosing and propagating plants that

MONOCULTURE OF WHEAT creates this land pattern in Wyoming. In monoculture the same crop is grown on the same land for at least two crop cycles. Here wheat fields lie side by side with fallow fields; in the next year the uses of the fields would be reversed.



STEAM TRACTOR began the trend toward mechanization on U.S. farms, and mechanization was one of the factors influencing the trend toward monoculture since 1900. The tractor served for plowing and also (as in this photograph) for threshing wheat.

MODERN COMBINES, harvesting wheat in Colorado, exemplify the increasing specialization of farm machinery in the U.S. In many cases the farmer's investment in specialized machinery is a strong incentive for him to shift to or continue in the monoculture (in this case wheat) that takes advantage of the machinery's particular capabilities and exploits technology familiar to the farmer.

exhibited superior or desirable traits. As Gregor Mendel's discoveries of genetic principles were rediscovered and became widely known late in the 19th century, the systematic crossing of plant varieties with desirable traits began in an effort to develop superior varieties. Theoretical and statistical principles of plant breeding were developed. These techniques gave rise to crop varieties resistant to or at least tolerant of many diseases and insect pests. Other desirable attributes of crops were improved also, including yield, grain quality and efficiency in the use of water. By concentrating on a single improved crop, the farmer could exploit its traits to the utmost.

Meanwhile the increased availability of land for crops other than livestock feed made it possible to introduce new crops. The prime example is soybeans, which came to the U.S. from China early in this century but for a long time were grown on a limited scale, mainly for hay. Today, reflecting the increased demand for vegetable oils, 35 million hectares are devoted to soybeans. Sunflowers are another new crop, grown on about 2.5 million hectares. A third new oilseed crop, canola, is grown on a million hectares in Canada. By specializing in such a crop the farmer could hope to maximize the farm's income.

Chemicals, the third technology underlying the shift toward monoculture, are typified by manufactured nitrogen fertilizer, which has become prominent since World War II. A parallel expansion of the agricultural-pesticide industry has taken place over the past 40 years. Because of these developments the farmer growing a single crop can have confidence that the fertilizers will make it flourish and the chemicals will protect it from pests and disease.

Socioeconomic Factors

In addition to technological changes as an influence on monoculture, several social and economic forces have been at work. For example, farmers were formerly pushed toward self-sufficiency because of limited financial resources and poor marketing and transportation facilities, but today the farmer buys almost all the goods and services utilized on the farm. The abandonment of the concept of total self-sufficiency means it is no longer imperative for each farm to diversify. Instead specialization toward monoculture is not only possible but also often dictated by economies of scale.

The enterprise of farming has changed in another way. With large increases in the amount of equipment and services to be bought and in the cash flow needed to run a farm, the farmer's call for credit has expanded greatly. In 1979 the average farm had production costs of $87,000, compared with $1,800 in 1910–14. Investments in machinery and supplies can be reduced somewhat by specializing in one crop or perhaps two. The farmer gains economy of scale and also avoids having to master the technologies needed to produce several crops at profitable levels.

Another influence is the farm policy of the Federal Government. Many Federal programs have manipulated farm income through controls on both production and prices. Government policies have essentially established the economic framework for farming, including net return to the producer. Unfortunately the Government's policy has not always been compatible with other goals of a good farming enterprise, often resulting in damage to soil and water resources, environmental degradation and disruption of social structures. The individual farmer, whose resources are limited, is forced to "farm the Government program"

DAMAGE TO LAND can result from monoculture, partly because the land is bare between crops and so is vulnerable to erosion (*top*). In a similar field (*bottom*) the problem was overcome by grading the gully and planting grass to reduce the erosive effect.

because the alternatives are seldom economically feasible.

At the same time most of the Government's programs provide some stability of farm income. Hence the farmer has a reasonable idea of how various cropping systems may affect his net income. Often such calculations lead him to the conclusion that monoculture or simple rotations provide the largest or stablest income.

Advantages and Limitations

A farmer who has turned toward monoculture can look for both financial and organizational benefits. Suppose he decides on a monoculture of wheat. Thereafter the only equipment he has to buy is machinery designed for work with wheat. He will use the machines on all his land, so that the land area covered per year per machine is maximized. This farming program usually results in the most efficient use of machinery, reducing equipment costs per unit of production to a minimum.

A monoculture enables the farmer to be a specialist in production techniques. If the crop is wheat, he can develop the management practices for fertilizer, pest control, tillage and so on that are best suited to his particular combination of soils, climate, labor supply and economic resources. If the farmer chooses to grow additional crops, he must apply different practices and must integrate the production requirements of all his crops into a workable and economically viable system. The task often requires compromises in such matters as herbicide mixtures and fertilizer that may result in lower crop yields.

The complexity of marketing is reduced with monoculture. The farmer need only keep up with one market—say the corn market. Moreover, the volume of product generated with monoculture enables him to take advantage of special markets or transportation discounts that may arise.

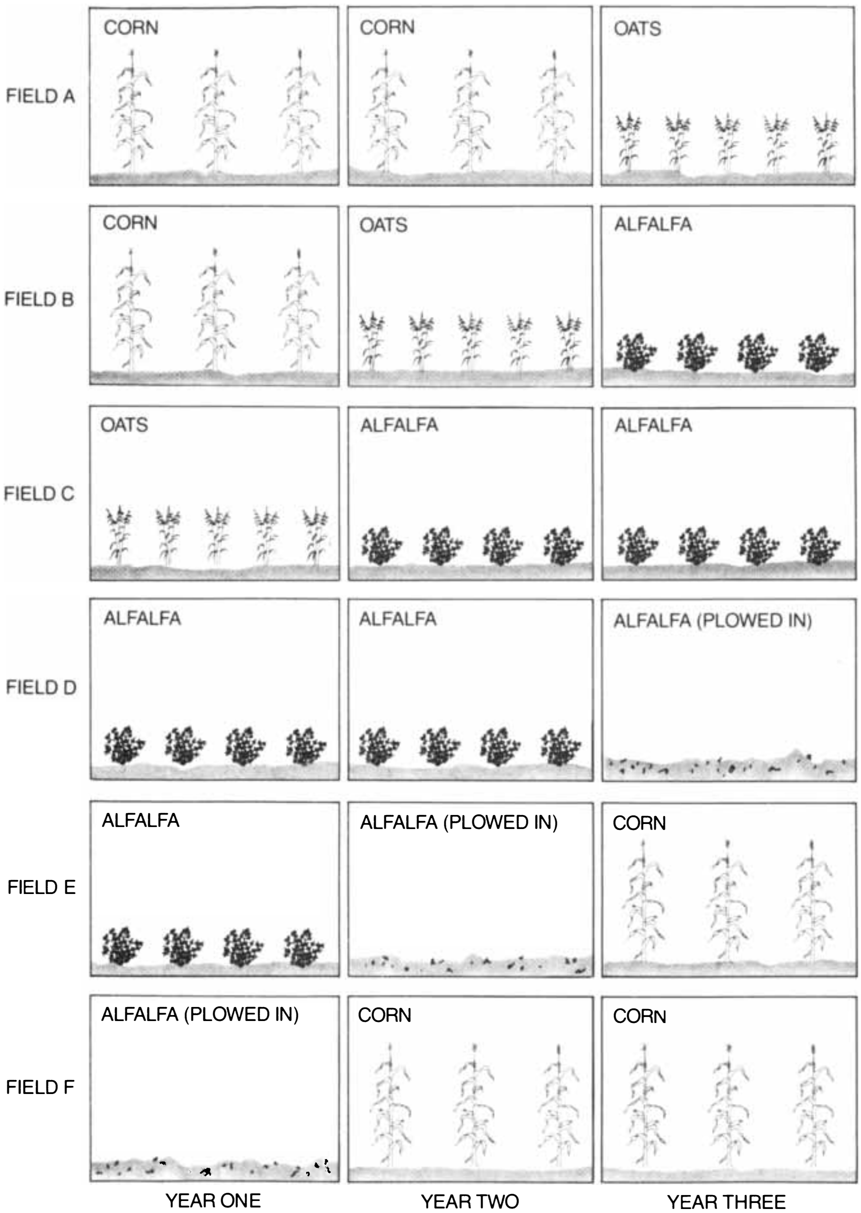
Farms specializing in monoculture seldom have livestock as an integral part of the enterprise, although associated activities such as cattle fattening and range operation may be included. The exclusion of livestock as an essential part of the business greatly simplifies farm management. Crop rotations are not necessary, the need for much fencing is eliminated and valuable time is not siphoned off during critical planting and harvesting periods to attend to the animals.

On the other hand, monoculture has certain limitations. One of them is that it puts the farmer at risk of economic instability. With one crop he is dependent on a single market and a particular set of economic conditions. This risk has been reduced by Federal support programs and by crop insurance, but it is still there.

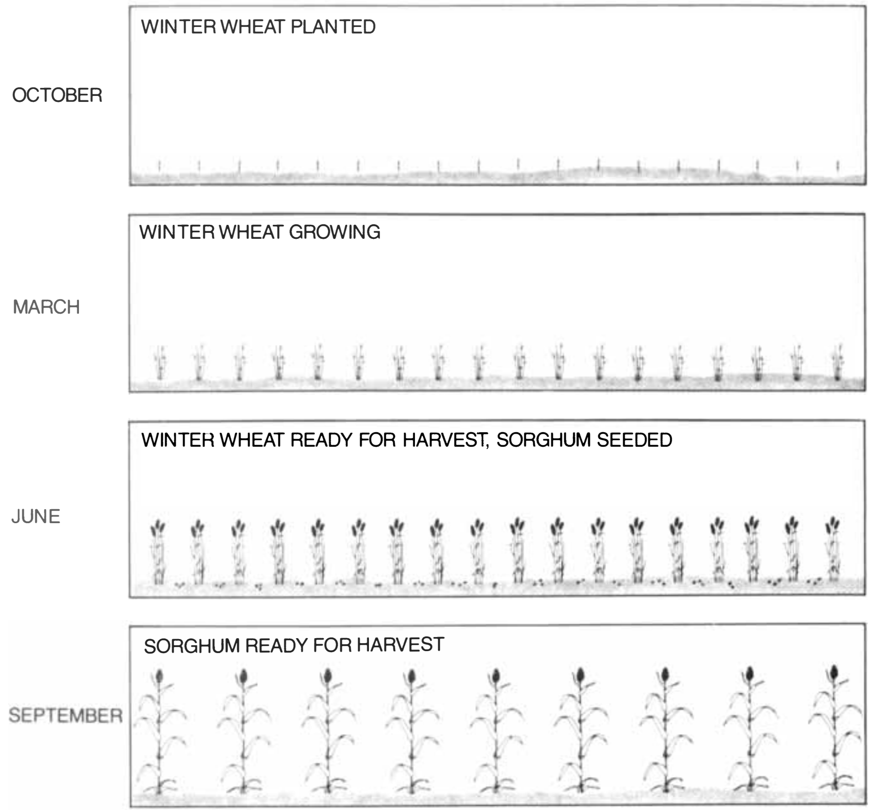
The accounting practices in agriculture may obscure the farmer's economic picture. Usually the economic and material inputs for each crop grown by a farmer are considered separately. The complementary relations among crops, such as a legume in rotation with another crop, seldom receive full economic recognition.

Where monoculture is dominant, a supporting economic and material infrastructure usually develops, reinforcing the position of the dominant crop. The tobacco and cotton economies of the Southeast up to about World War II typify the situation. Few of the people caught up in that economy foresaw its devastating consequences: the degradation of soil resources, particularly the loss of a major part of the topsoil in the southern Piedmont Region, and the resulting disruption of the economy.

The experience in the Southeast brought home the fact that erosion is usually accelerated by cultivation and is often likely to be severest when a



SIX-YEAR CROP ROTATION is an alternative to monoculture. It calls for six fields or groups of fields. A typical six-field rotation is shown here for its first three years. For the entire six-year program the pattern of rotation for each field is corn, corn, oats, alfalfa, alfalfa, alfalfa. In the sixth year the alfalfa is plowed under to improve the fertility of the soil. The rotation system has the advantage of drawing different nutrients from a field and, with the inclusion of alfalfa (a legume), also of increasing the field's nitrogen store.



DOUBLE CROPPING provides a means of intensifying the output of a farm field while still achieving the effect of rotation. In this version two different crops are harvested from one field in the course of the same year. The second crop (here sorghum) is planted just before the first crop (winter wheat in this example) is harvested in the spring.

single crop is grown repetitively. Multiple cropping can protect the soil, thereby maintaining its productivity, by increasing the amount of vegetative cover. An excellent way to restore productivity to eroded soil is to begin with a legume cover crop or a rotation system, either of which will simultaneously reduce erosion and restore organic matter to the soil.

Erosion, Rotation and Pests

Erosion changes the hydrologic and physical properties of the soil. It decreases fertility by carrying away nutrients and by exposing subsoil that is less fertile and may be acidic. The lost nutrients include nitrogen, phosphorus and potassium; in humid regions particularly the availability of certain micronutrients is also reduced. About 60 kilograms of nitrogen is lost with each metric ton of organic matter removed by erosion. Crop yields are usually lower on eroded soil, no matter how much fertilizer is applied. The loss of organic matter also increases the risk of injury to the crop by herbicides. Organic matter absorbs or inactivates herbicides applied to the soil; when it is gone, the ordinary rate of application of herbicides may be excessive.

The sequential order in which crops are grown often markedly affects the yield of the following crop, a phenomenon known as the rotation effect. For example, alternating corn with either a legume or a nonlegume crop normally improves the yield of corn over the yield obtained in monoculture. The rotation effect is not fully understood; it could be the result of improved soil fertility, pest and disease control, the quantity and quality of organic matter, water conservation and perhaps other factors. Crop rotations often provide a more continuous cover for the soil than monoculture, and they also help to control erosion and such pests as weeds, disease and insects.

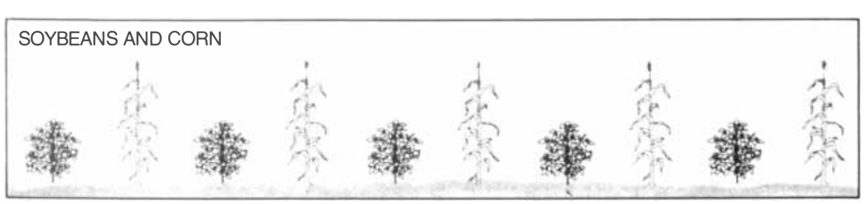
From the earliest days of agriculture it has been known that monoculture often results eventually in decreased yields, even with supplementary fertilization, and that certain crops have a beneficial effect on the growth and yield of the crops that follow. A particularly useful technique is to grow "green manure," a crop that is plowed into the soil to improve its condition and to provide nutrients for the next crop. Legumes are highly desirable for the purpose because of their nitrogen content. There are many combinations of crop sequences, tillage, fertilization, fallowing and other procedures that result in a beneficial rotation effect on the next crop.

Crop rotation also serves as a means of controlling pests. Its success at the task comes from the fact that continuous cropping forms ecological niches where pests can get a strong foothold. Their life cycle can be interrupted by inserting different crops or by leaving a field fallow. The technique works in various ways. It may simply starve out a particular species of pest. Sometimes the new crop itself acts through its own chemicals to deter certain pests. Or the new crop may enhance biological control by pitting new pests against the resident ones.

The hazards of monoculture can extend beyond a single farm to include an entire crop. In many cases a major crop such as corn is dependent on only a few original sources of germ plasm, or hereditary material. The genetic base is narrow, making the crop vulnerable to a new or newly virulent disease. The U.S. had a scare of this kind in 1970 when a mutant form of Southern leaf blight (the fungus *Helminthosporium maydis*) destroyed about 10 percent of the corn crop.

Beyond the specific limitations of monoculture is a general one: the region or the farm practicing monoculture is frequently not achieving the maximum agricultural potential offered by the land and the climate. A number of alternative cropping techniques may utilize these resources more efficiently. The alternatives include fixed rotation, multiple cropping and intercropping.

A fixed rotation, the commonest alternative, is a system in which crops



INTERCROPPING, another form of multiple cropping, entails growing two crops simultaneously in one field. It too is an alternative to monoculture. Here the crops are soybeans and corn in adjacent rows. Other options are orchard-legume and grass-legume.

are grown in an unchanging sequence. A fixed rotation widely practiced in the U.S. is that of corn (one crop or more) followed by soybeans. Although the soybean is a legume and therefore contributes nitrogen to the soil, it leaves less residue than corn does, and so this simple but intensive rotation may result in more erosion on some soils than would take place with a monoculture of continuous corn. One solution is to include a forage crop such as a grass in the rotation. It not only protects the soil while it is growing but also reduces erosion while the next crop is growing. The solution is particularly attractive when livestock are at hand to utilize the grass.

The possibilities in rotation include row crops, small grains, legumes, grasses, legume-grass mixtures and leaving a field fallow. Both diversity and relatively constant production of each crop can be achieved by dividing the land into as many parts as there are years in the rotation. A six-year rotation would require six fields or groups of fields of roughly equal size and productive capacity. An example of a six-year rotation might be corn, corn, oats, alfalfa, alfalfa, alfalfa, with the alfalfa plowed under in the final year as a green manure.

Rotations do not have to be fixed, and in some conditions it is better to make them variable. Ardell D. Halvorson, then at the U.S. Department of Agriculture's Northern Plains Soil and Water Research Center in Sidney, Mont., and several colleagues designed a computer model called FLEXCROP. FLEXCROP bases crop selection on the preceding crop and on the amount of water stored in the soil and available to the plants. This model and variations of it are in service in many parts of the U.S. and Canada. For the conditions in Montana the program lists winter wheat, spring wheat, barley, oats, safflower and fallow as alternatives for rotation. A typical rotation might be winter wheat, spring wheat and summer fallow in place of the two-year sequence of winter wheat–fallow that has been traditional in the Great Plains. The new sequence results in two wheat crops in three years rather than two in four years.

Another alternative to monoculture, and also a type of rotation, is the planting of a seasonal cover crop. Such a crop is grown to protect the soil rather than to be harvested; it fills periods when the ground would otherwise be bare. Because such periods are usually cold or dry, the plants have to be hardy and easily established. The possibilities include a number of clovers (sweet, red, crimson and arrowleaf), vetches, Austrian winter peas, rye, oats and ryegrass. The effectiveness of the seasonal cover crop in controlling erosion can be enhanced by leaving the residue of the crop on the soil surface when planting the next crop.

Multiple Cropping

Multiple cropping entails growing two crops or more on the same field in a single year. It intensifies the farm's production in time or space. A method that takes advantage of time is sequential cropping: growing of crops one after the other. The second crop is planted just before or after the first one is harvested. A method that takes advantage of both time and space is intercropping, which means growing two crops or more simultaneously on the same field. The benefits include increasing the total production per unit of land and greater dependability of return compared with monoculture.

Growing two crops in sequence in the same year is an ancient practice that began to gain acceptance in the U.S. after World War II. The trend has been encouraged by recent shifts to methods of tilling that seek to conserve the soil, by the development of plant varieties that mature early, by increases in supplemental irrigation and by better farm management. An experimental technique being tried in irrigated areas of the central U.S. entails adapting a center-pivot irrigation machine, which in normal use distributes water in a broad circle, to spread corn or sorghum seeds before winter wheat is harvested and similarly to seed winter wheat before the corn or sorghum is harvested.

A common pattern of double cropping in the South and in the southern part of the Corn Belt is fall-seeded grain followed by soybeans. The pattern of wheat followed by soybeans,

```
******* ECONOMIC ANALYSIS *******

WHEAT

VARIABLE COSTS PER ACRE:

SEED                               6.00
NITROGEN                           0.00
PHOSOPHORUS                        6.90
WEED CONTROL                       5.00
MACHINERY                          7.56
LABOR                              4.00
------------                      -----
TOTAL                             29.46

FIXED COSTS PER ACRE (EXCEPT LAND):

TAXES                              2.00
INSURANCE                          3.00
MACHINERY OWNERSHIP               15.00
RETURN TO MANAGEMENT               5.40
--------------------              -----
TOTAL                             25.40

GROSS RETURN PER ACRE             90.00
RETURN OVER VARIABLE COSTS        60.54
RETURN OVER VARIABLE
   AND FIXED COSTS                35.14
```

FLEXCROP PROGRAM is a computer model designed to help a farmer decide whether an alternative to monoculture would suit his needs. The program asks him questions about the condition of the soil, the amount of water available and his farming practices. In this case the farmer decided to stay with a spring-wheat monoculture and asked for an economic analysis; the computer program provides it here in dollars per acre per year.

limited to the Southern states before the adoption of no-tillage farming, has now spread to Ohio, Illinois and Indiana. Farther north the double cropping of grain and row crops, often harvested as silage, is fairly common. Another practice in several states of the Corn Belt is to grow canning peas under contract with major canneries, harvest the peas in the spring and follow them with early corn or soybeans. Although the practice of double cropping is increasing, the total area devoted to it is still fairly small. In some regions, however, notably Appalachia, the Southeast and the Mississippi delta, a substantial amount of land is farmed in this way.

No-tillage farming increases the potential for multiple cropping. What has made the no-tillage practice possible is the development of herbicides that control weeds well enough to eliminate the need for plowing and tilling. With no plowing to do the farmer can plant his second crop sooner. Because a mulch remains on the soil, the water loss associated with plowing is eliminated and the new crop has a better supply of water. With these advantages the farmer can go to multiple cropping and make more intensive use of his machinery.

The third alternative to monoculture, intercropping, entails growing two or more crops on the same land at the same time. A typical pattern is an orchard combined with a grass or a legume that grows on the orchard floor. Another common practice in intercropping is a mixture of a legume and a grass; they serve as pasturage or for making hay. The practice of seeding soybeans into wheat before the wheat is harvested also constitutes intercropping. In the Southeast it is possible to grow two row crops simultaneously. Intercropping brings about a variety of complex interactions between the crops because they both share and compete for nutrients, water, light, oxygen and carbon dioxide.

Measures of Success

Two measures of the degree of success of intercropped systems are the land-equivalent ratio and the income-equivalent ratio, usually abbreviated to LER and IER. The LER compares the area required for a single crop with the area required for intercropping to obtain an equivalent yield. The IER compares the area needed for a single crop with the intercrop area that generates the same gross income, given the same level of land management; it represents the conversion of the LER into economic terms.

An example of a highly successful

system of intercropping is legume-nonlegume. The legume fixes much of its own nitrogen, and the nonlegume exploits the mineral nitrogen (nitrates and ammonium) present in the soil. The practice improves the LER on soils deficient in available nitrogen. A typical pattern in the West and Middle West is to plant a grain as a nurse crop when a new stand of alfalfa or some other legume is being established. Growing yellow sweet clover along with the grass on rangeland can be successful in some years on the northern Great Plains as a means of increasing the production of forage. Alternating rows or groups of rows of soybeans and corn has shown an advantage in yield, although harvesting can present problems.

Nonlegume intercropping arrangements are also possible. An example in the Southeast is the intercropping of tall fescue with a sorghum–Sudan grass hybrid and rye. The average yield of forage was 150 percent higher than it was when tall fescue was grown as a monoculture.

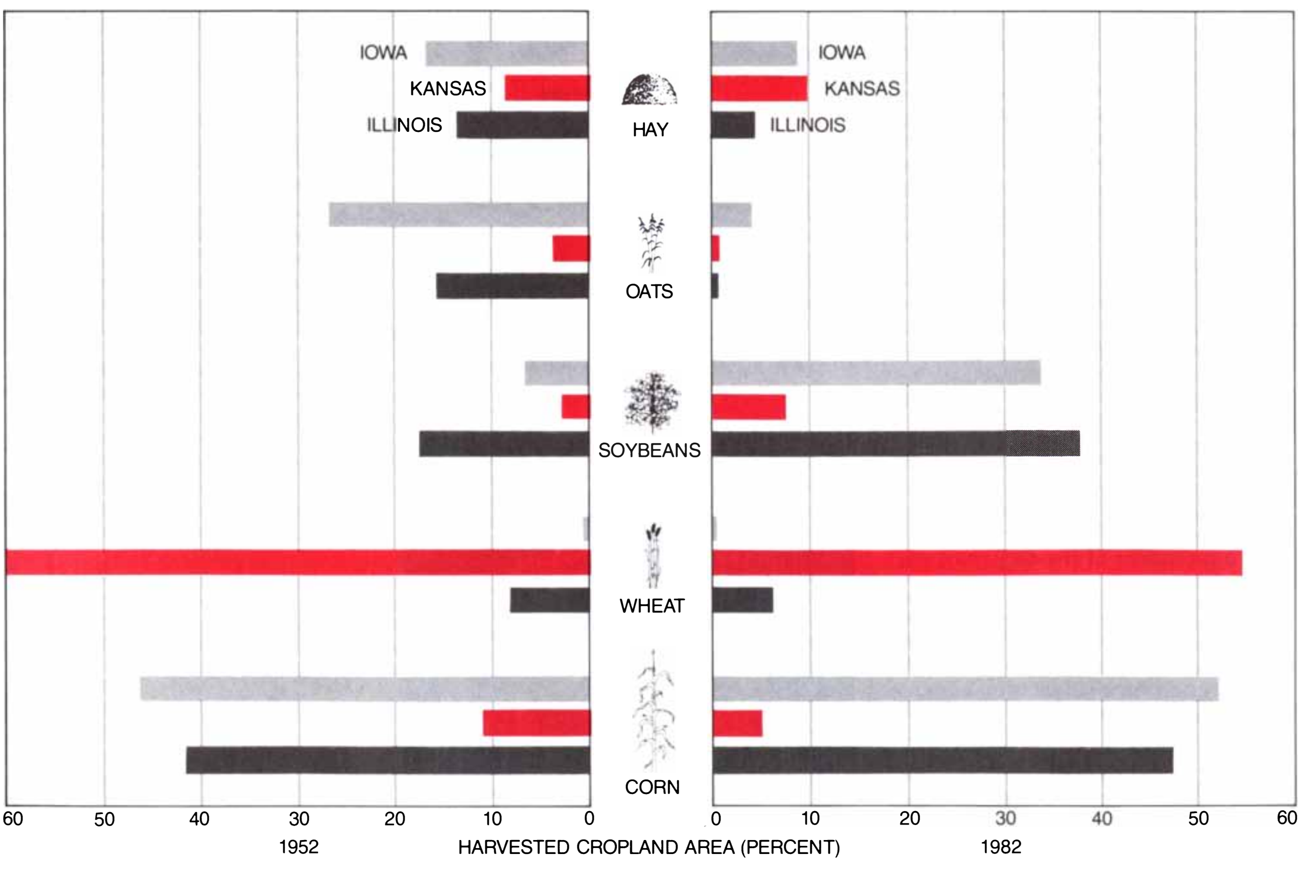
There is evidence that intercropping can suppress pests. An intercrop consisting of corn and peanuts has been reported to decrease the impact of the corn borer. Moreover, an intercrop system that maintains a constant cover can reduce weeds by intercepting much of their light.

Future of Monoculture

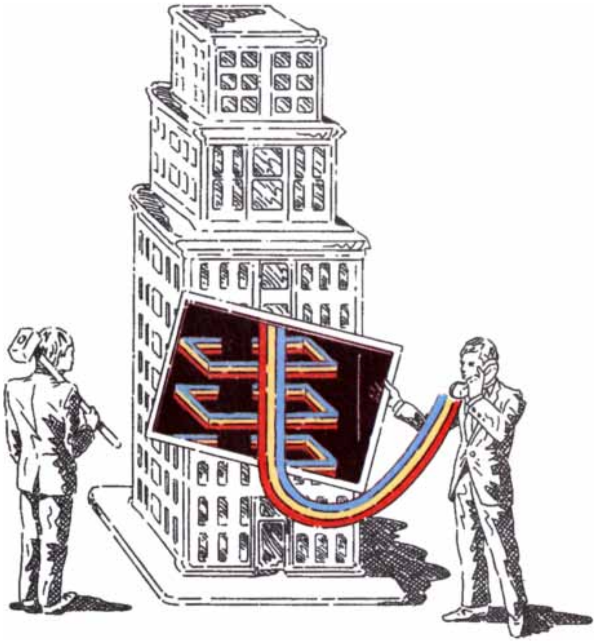
The driving force behind a farmer's choice of production techniques is usually the net economic return. For several decades a number of factors have combined to make the immediate return of monoculture look attractive. The situation may be changing, however. For one thing, the energy crises of the 1970's (presumably not the last of their kind) have increased the cost of the farmer's inputs at a time when the value of his outputs has been declining. The economic advantage of specialized monoculture farming has therefore diminished.

In addition, the environmental effects of monoculture can be expected to influence the long-term economics of the practice, and those effects are coming to be better understood. Monoculture often tends to degrade the environment by accelerating soil erosion, increasing the potential for depleting or degrading groundwater resources, reducing the quality of surface water and using up fossil energy resources. Many of these effects can be controlled within tolerable limits by good management practices. They entail a cost, however, from which the farmer seldom receives an immediate economic return.

Because the conservation of resources benefits all segments of society, Government programs are sometimes developed with the intent of passing on part of the cost of conservation programs to the society as a whole. Many other farm programs have encouraged monoculture farming without providing incentives for environmental protection. It seems likely that Government programs will gradually do more to promote soil and water conservation and that farmers will move increasingly toward no-tillage agriculture and other practices that aid conservation. Both these trends would encourage a shift toward alternatives to monoculture. It can be expected that improved computer models will identify economically and environmentally acceptable alternatives to monoculture, making the shift easier for farmers to undertake. Further research will lead to better cropping alternatives and management practices, so that the farmer can simultaneously optimize his income and protect the environment.



SHIFT IN CROPS between 1952 and 1982 is shown for three major farm states. Hay and oats are down because of the decline in the number of farm animals in grain-growing states. Soybeans have taken over much of the farmland once devoted to feed crops.



Before you break down walls to rewire for a LAN, look again.

You can use your existing phone lines for a LAN if you have a Centrex telecommunications system from New York Telephone.

Until now, even the most sophisticated Local Area Network (LAN) was installed through the most unsophisticated means. It meant breaking walls, drilling holes, fishing for wires, and inconvenience.

But today, the process of getting a LAN has finally become as advanced as the LAN itself.

Now, if you have a Centrex telecommunications system, LAN capability can be added quickly and easily using your existing telephone lines...without rewiring, wall bashing, or even downtime.

Once you start using the Centrex LAN, more advantages appear. For example, it integrates equipment from different manufacturers. Personal computers, terminals and phones from different makers can work together, communicating at speeds up to 19.2 kbps asynchronous service.

And since it's central-office-based, we do the maintenance of the Centrex LAN from our premises. More capacity can be switched on whenever you need it. And initiating service requires no capital investment.

To find out more about a Centrex LAN or the Centrex telecommunications system, call New York Telephone at **1 800 942-1212, ext. 44,** or one of our authorized sales agents.

New York works on New York Telephone.

New York Telephone

A NYNEX Company

Authorized Sales Agents:

AIM
Telephones Inc.
1 212 777-6600

ALL-Mode
Communications, Inc.
1 607 347-4164

Centel Financial
Systems
1 212 514-7800

Computer Telephone
Corporation
1 800 225-3288

CONTEL IPC
1 212 825-9060

DIAL
Communications, Inc.
1 212 868-6677

Digital Support
Systems, Inc.
1 212 777-3333

Executone
Albany, Inc.
1 518 869-3571

Executone
Long Island
1 516 582-3200

Executone
Telecom, Inc.
1 716 288-2000

Executone
of Hudson Valley
1 914 564-4880

Gaffney
Communications, Inc.
1 315 797-8222

NYNEX
Business Information
Systems Company
1 800 346-9X9X

Ronco Communications
& Electronics, Inc.
1 716 879-8100

Standard
Telecommunications
Systems, Inc.
1 212 532-5440

Tel Plus, Inc.
1 800 TEL-PLUS

Twin State CCS
1 518 563-7100

Worldwide
Communications Inc.
1 718 361-1070

Optical Neural Computers

Can computers be built to solve problems, such as recognizing patterns, that entail memorizing all possible solutions? The key may be to arrange optical elements in the same way as neurons are arranged in the brain

by Yaser S. Abu-Mostafa and Demetri Psaltis

Computer scientists find it increasingly frustrating to see how casually a three-year-old picks out a tree in a picture. Sophisticated programs running on the most powerful supercomputers are capable of only a mediocre performance in doing what essentially amounts to the same task: pattern recognition. What makes this state of affairs so paradoxical is the fact that solutions to many problems that overtax the human brain can be arrived at quickly by computers. Indeed, a simple pocket calculator can easily outperform the human brain in such tasks as finding the product of two 10-digit numbers. What is the difference between the multiplication of numbers and the recognition of objects that makes the latter so much tougher to achieve in computers? In other words, why is it so difficult to make a computer recognize a tree?

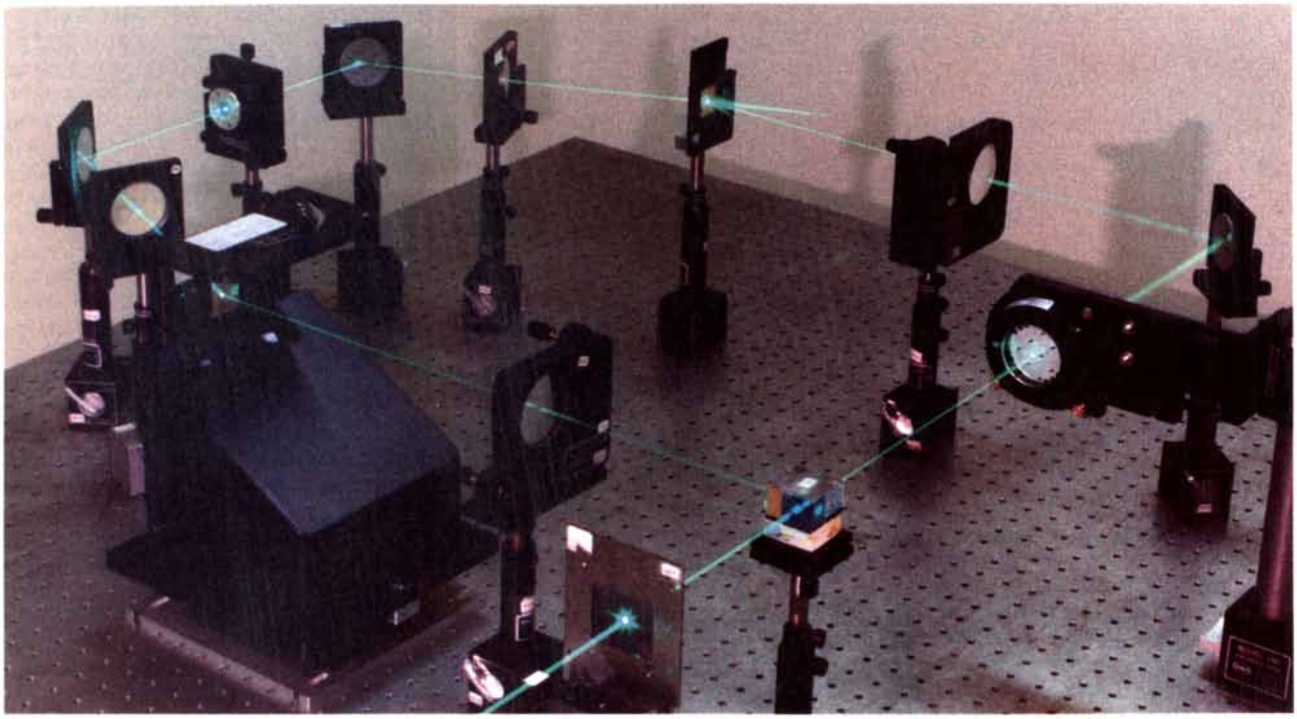
The answers to these questions ultimately hinge on the fact that pattern-recognition problems cannot be compactly defined. In order to recognize trees a comprehensive definition of a tree is required, and such a definition would be tantamount to a description of every conceivable variant. Problems such as those posed by pattern-recognition tasks constitute a subset of what we call random problems: problems whose solution requires knowledge of essentially every possible state of a system. Solving a random problem therefore entails memorizing the set of all possible solutions and quickly selecting the best solution from the set, given the input data. In contrast, the solution to such a classical computation problem as multiplication can typically be expressed succinctly in terms of an algorithm: a sequence of precise instructions specifying how the input data are to be manipulated to arrive at the solution.

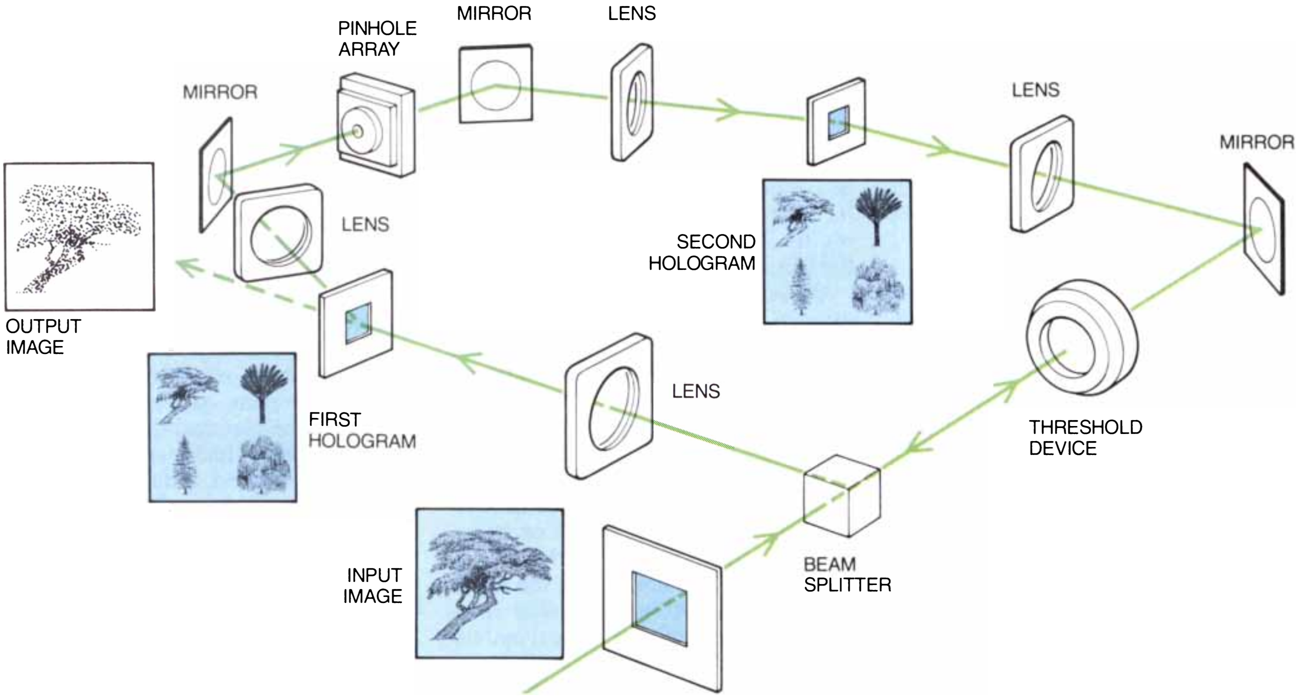
A conventional computer is adept at mechanically executing the instructions in an algorithm, but it cannot match the memorization and recollection capability of the human brain, which regularly and effortlessly conquers pattern-recognition problems. Because the brain is unique in its capability to solve random problems, many computer scientists and mathematicians have taken a closer look at how the brain works in the hope that the principles of its operation can be fruitfully applied in designing machines capable of solving random problems. Devices designed to model the workings of the human brain by emulating its anatomic structure are called neural computers; like the brain, they would consist of a large number of simple processors that are extensively interconnected. In this respect one technology stands out as being particularly promising for constructing neural computers: optics.

Optical technology dovetails nicely with the notion of a neural computer because the technology's strengths lie in exactly those areas that distinguish a neural computer, such as the interconnection of a large number of processing elements; its weaknesses lie in areas that are less critical for the functioning of a neural computer, such as the ability to perform intricate logic operations at the processor level. Whereas semiconductor technology in conventional computers has proved to be capable of tackling classical computation problems by means of algorithms, optical technology in the neural computers we envision may one day make it possible to solve random problems efficiently. Indeed, in our laboratory at the California Institute of Technology we and our colleagues have already built experimental pattern-recognition systems that represent a first step toward an optical neural computer.

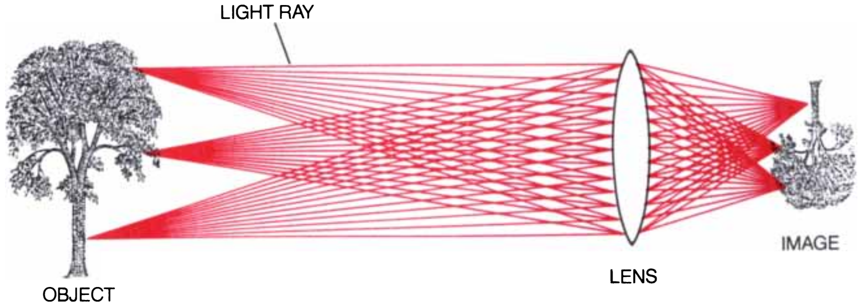
Regardless of the technology a computer incorporates (be it optical or electronic) or the functions it executes (be they multiplication or pattern recognition), two principal activities take place in it as it solves a problem: logic operations and data transmission. Viewing computation in such fundamental terms helps to get to the source of a particular computer technology's strengths and weaknesses. Semiconductor technology can be applied to build sophisticated logic circuits from electronic switches of very small size that have very reliable characteristics, yet such integrated circuits are rather limited in the amount of data that can be transmitted among the circuit elements. The reason is that on a silicon chip communication links consist of wires that must be kept separated by at least a minimum critical distance; otherwise the electrical signals they carry interfere with one another. This practical restriction places an effective limit on the number of wires that can be placed on a chip and hence on the amount of data communication that can take place on the chip.

Is there another technology from which computers could be built that does not suffer from this limitation in data communication? The operation of the eye's lens suggests one. The lens takes light from each of millions of points in the entrance pupil of the lens and redistributes it to millions of sensors in the retina. It is in this sense that the lens can be thought of as a highly capable interconnection device: light from every point at the pupil is "connected" to every point in the image focused on the retina. Moreover, multiple beams of light can pass through lenses or prisms and still remain separate. Indeed, two beams of light, unlike a pair of current-carrying wires, can cross without affecting each other. It is the ability to establish an extensive communication network among processing elements that primarily distinguishes optical technology from semiconductor technology in its application to computation.





PATTERN-RECOGNITION SYSTEM (*top*) **developed by the authors and their colleagues at the California Institute of Technology can quickly find the best match between an input image and a set of holographic images that represents its "memory." The input image is projected into the system (*diagram*) through a beam splitter—a partially reflecting mirror—by illuminating a transparency (in this case one carrying an image of a cypress tree) with a laser beam (*bottom left*). The light that passes through the beam splitter hits the front of a threshold device, reflects off it and retraces its path back to the beam splitter, where it is reflected at an angle to initiate an optical "loop." A lens focuses the input image on a hologram, where the image interacts with each of four holographically stored images (here, of trees), creating patterns of light whose brightness varies according to how well the input and stored images match. A lens and a mirror direct the light issuing from the hologram to a pinhole array that spatially separates the four light patterns associated with each combination of input image and stored image. Another lens and mirror collimate the light and illuminate a second hologram with it. This hologram contains the same set of stored images as the first and is designed to produce a superposition of the four image combinations. The beam bearing the superposed images is focused by a third lens-and-mirror pair on the back of the threshold device. The pattern of light impinging on the back of the threshold device determines what light is reflected off its front. Since the brightest image reaching the back of the threshold device represents the best match to the input image from the set of stored images, it is essentially an image of the best match that is reflected from the front of the device into the optical loop for a second pass. Successive passes around the loop continue to enhance the best match from the set of stored images, which can ultimately be retrieved as the output image leaving the system through the first hologram.**



OPTICAL LENS is an inherently powerful interconnection device: it connects every light ray that originates at a point on an object and passes through the lens with every point of the object's image. Unlike wires on an integrated-circuit chip, light rays can come close to one another and even cross without affecting one another. Hence millions of light rays could conceivably carry data simultaneously into a processing device, whereas electronic devices on a chip are limited to accepting input from a few wires at a time.

Because optical processing elements communicate through beams of light, they can be hooked up to one another without attaching a cumbersome wire between each pair of elements, and they need not be confined to the restrictive planar configurations of silicon chips. Indeed, optical connections are being considered as a means of relieving communication bottlenecks encountered in very-large-scale-integration chips. In such a hybrid optoelectronic system the processing units are electronic but the connections between them are optical, typically consisting of light sources and light detectors fabricated on the same chip as the processing units.

The most promising device for establishing arbitrary optical connections is not a lens but a hologram. Holograms are best known as a means of generating three-dimensional images, but more generally they represent an effective technique for recording and reconstructing the intensity of a light ray as well as the direction from which it came. Whereas a conventional lens maps each light ray entering the lens to a particular point on the image plane, holograms can readily be "programmed" to allow a variety of such mappings.

A planar hologram, produced on a relatively thin medium such as photographic film, can direct any light beam on one side of it to any point on the other side, provided the total number of points and light beams does not exceed the number of resolvable spots on the film. The number of resolvable spots in a one-inch-square hologram can be as high as 100 million. This would allow each of 10,000 light sources to be fully interconnected with each of 10,000 light sensors. A similar interconnection scheme by means of wires would be extremely difficult to accomplish on a silicon chip.

Even more prodigious in its capability to connect light emitters to light detectors is a volume hologram made from a photorefractive crystal. When such a crystal is exposed to light, electric charges are generated in it that redistribute themselves according to the pattern of the illumination's intensity. Because the local charge density in a photorefractive crystal determines the local refractive index (a measure of how fast light travels through the material), holographic images projected onto the crystal are recorded in terms of the spatially varying refractive index. The image information can then be extracted from the hologram simply by illuminating the crystal with a light beam.

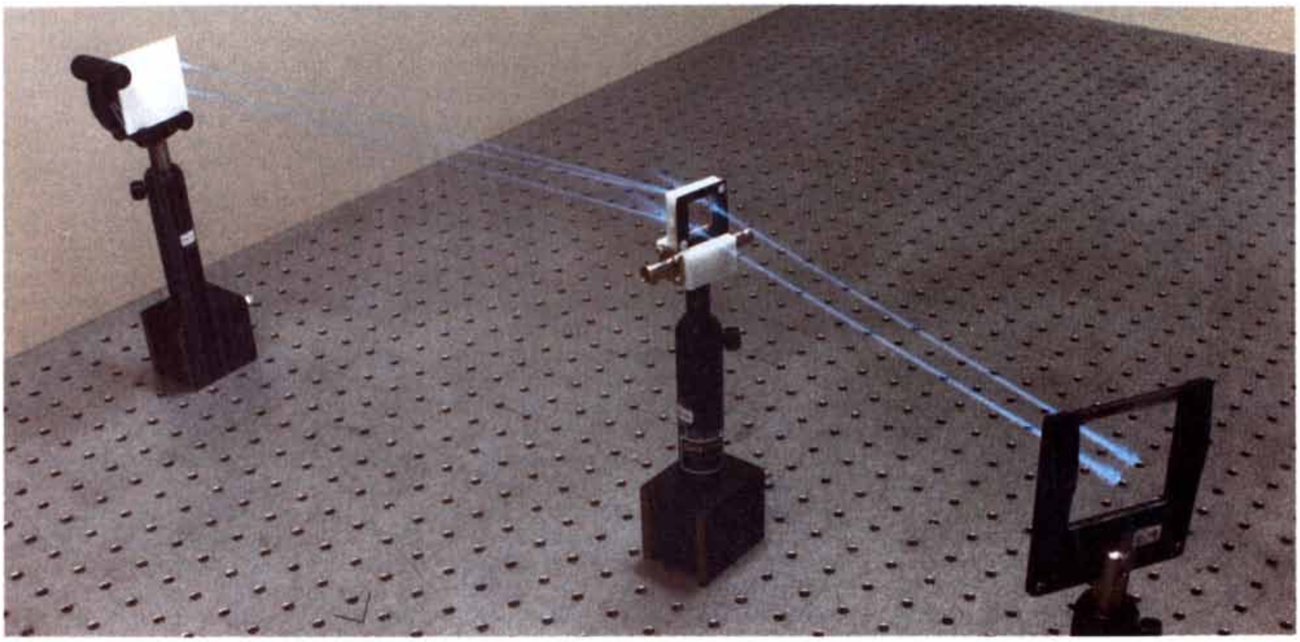
Other hardware associated with traditional computation can also be realized optically, namely switching elements (from which processors are constructed) and memory elements (in which data are stored). Switching elements can be made from a nonlinear optical material. An optical material is nonlinear if its transmittance properties, such as its opaqueness or its refractive index, change as the brightness of the light incident on the medium changes. Gallium arsenide is an example of a nonlinear optical material from which two-dimensional arrays of optical switches have been fabricated. Nonlinear optical materials make possible the construction of an "optical transistor," in which the brightness of one light beam controls the transmission of another light beam.

An optical memory element is essentially a device that can alter an input light beam into one of two possible states, each state corresponding to a bit of binary data (either a 1 or a 0). Optical memories have been developed for audio and video recording and more recently as digital mass memories for electronic computers. Yet in these devices the stored information is typically accessed serially by focusing a light beam on one stored bit of information at a time, much as information is read off a magnetic tape. These devices do not exploit the huge potential for increasing the speed with which data can be transferred from memory by allowing parallel access to stored data. Millions of bits of information could be read out and transferred at the same time merely by shining an unfocused light beam on a suitably designed optical memory device [*see illustration on page 92*].

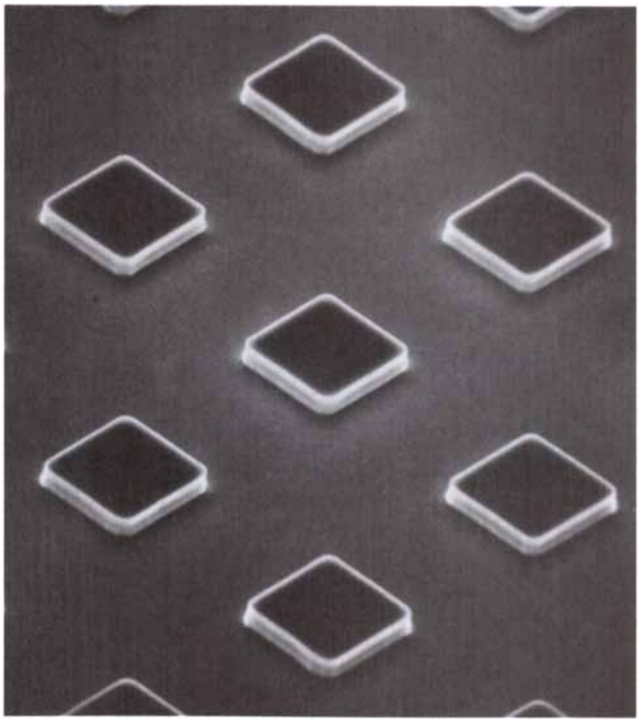
The fact that designers of optical memories have not exploited the potential for parallel access to data is an indication that most of the work in the development of optical switching and memory elements is done with the goal of implementing these devices in the execution of sequential, binary-logic functions. Hence these optical components would essentially duplicate (albeit perhaps more efficiently) the same operations that take place in conventional electronic computers. Although increased switching speed and a massive memory may ultimately result from such development efforts, they do not fundamentally change the mode of computation of conventional computers. Consequently devices in which electronic switches and memory elements have simply been replaced by optical analogues are just as likely as current computers to falter when they are confronted with pattern-recognition problems.

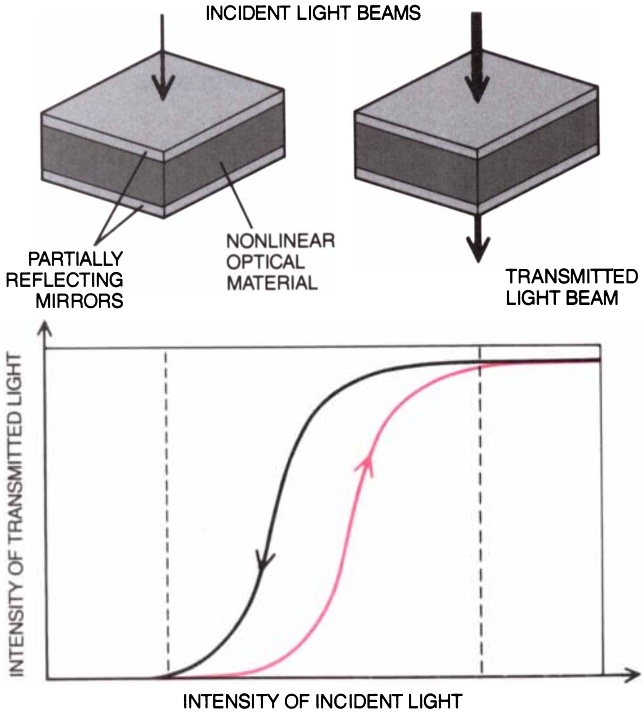
In order to understand why this is so it is necessary to consider how a conventional electronic computer solves a problem. As we have indicated above, the classical theory of computation, from which current computers were developed, is built around the notion of algorithms. The procedure for long division of two numbers is a good example of an algorithm. The procedure can be specified easily, and once it has been mastered—whether by a computer or by a sixth grader—it is universally applicable: it works as well for dividing a four-digit number by a three-digit one as it does for dividing a 1,000-digit number by a 900-digit one (although the algorithm may take longer to complete in the latter case, particularly for the grade schooler).

Computational problems that lend themselves to algorithmic solutions share a characteristic property: they are structured, meaning they can be stated clearly and concisely in mathematical terms. Most of the problems currently being solved with computers belong to this class of structured problems, and it is now a universal practice for computer programmers to look for



VOLUME HOLOGRAM (*on middle stand*), like the more familiar planar hologram recorded on photographic film, can distribute beams of laser light in "programmable" directions. Here, for example, two input laser beams coming from the bottom right are turned into four output beams heading toward the top left. A volume hologram is made from a photorefractive crystal. When such a crystal is exposed to light, electric charges are generated in it that redistribute themselves according to the pattern of the illumination's intensity. Because the local charge density in a photorefractive crystal determines the local refractive index (a measure of how fast light travels through the material), the crystal can record holographic images in terms of the spatially varying refractive index. Such a hologram can set up a pattern of optical connections between light sources and detectors for each image stored in it.





OPTICAL SWITCHING ELEMENTS (*left*) are manufactured by sandwiching a nonlinear optical material (which alters its refractive index according to the intensity of the light to which it is exposed) between two partially reflecting mirrors. An element thus constructed (*top right*) can abruptly change its transmission properties depending on the intensity of the incident light beam. It also exhibits a so-called hysteresis cycle (*bottom right*) when it is switching. As the intensity of an incident light beam is gradually increased from zero (*color*), the element does not allow any light through until the incident beam reaches a certain threshold intensity; then transmission quickly rises to a maximum value. The intensity of the transmitted beam will not retrace the same path if the intensity of the incident beam is reduced back to zero. Instead the abrupt change in transmission (*black*) now occurs at a lower incident-beam intensity. The photomicrograph was provided by Thirumalai Venkatesan of Bell Communication Research, Inc.

an algorithm whenever they have to solve a problem.

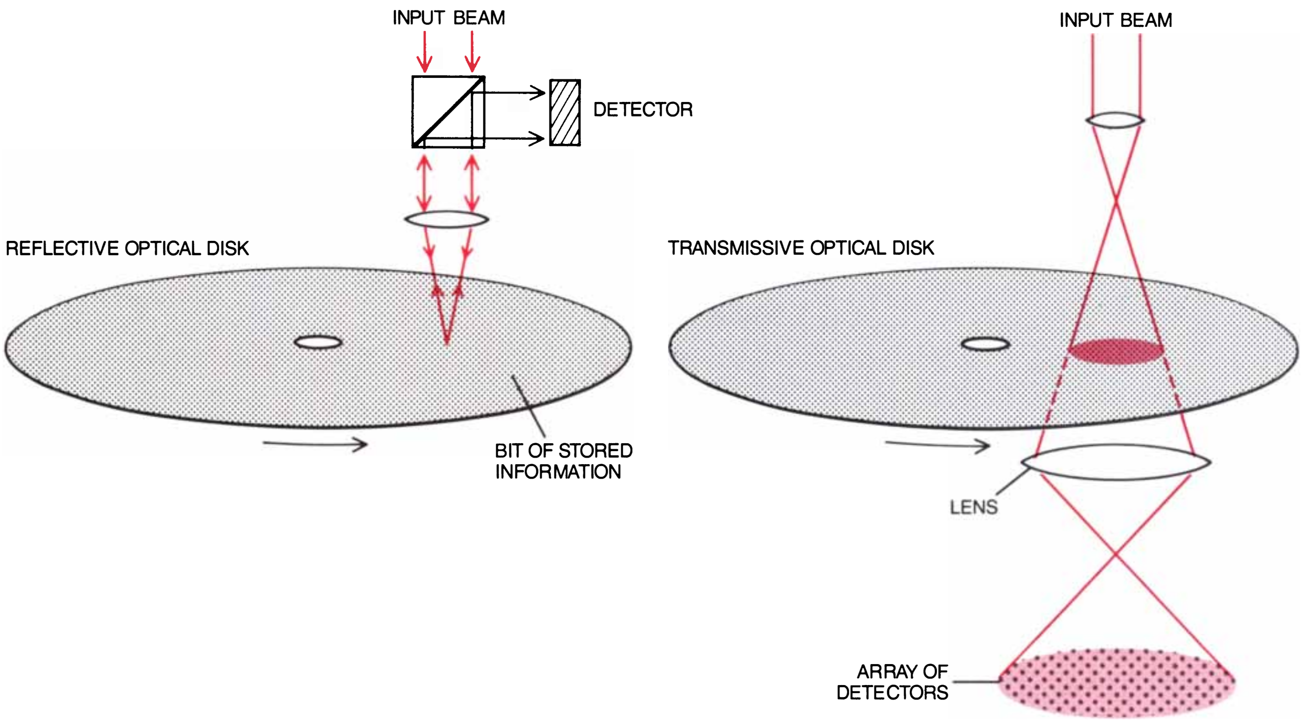
Problems such as pattern recognition in natural environments, however, lack the structure that would allow simple algorithmic solutions. It is this departure from the properties of structured problems and the methods for solving them that characterizes a random problem. The term "random," as we apply it here, is derived from the mathematical concept of randomness, namely the lack of a concise and complete definition. Randomness in this sense is linked to the mathematical notion of entropy, which can be thought of as the amount of disorder in a problem or, equivalently, the amount of information needed to define the problem. Because a formal description of a random problem would amount to a listing of essentially every possible solution to the problem, random problems have a much higher degree of entropy than structured problems.

To better understand what it means for a problem to be random, consider once again our tree-recognition example. Although it is clear to most people what a tree is, it would be very difficult to write down a concise definition for a visitor from another planet, who does not know what "branches" or "leaves" are, or for that matter what the color "green" is. Even if examples of each of these features could be presented to the alien visitor, there are innumerable types of branches, leaves and hues of green; a handful of examples is unlikely to be enough to cover all possible arboreal combinations.

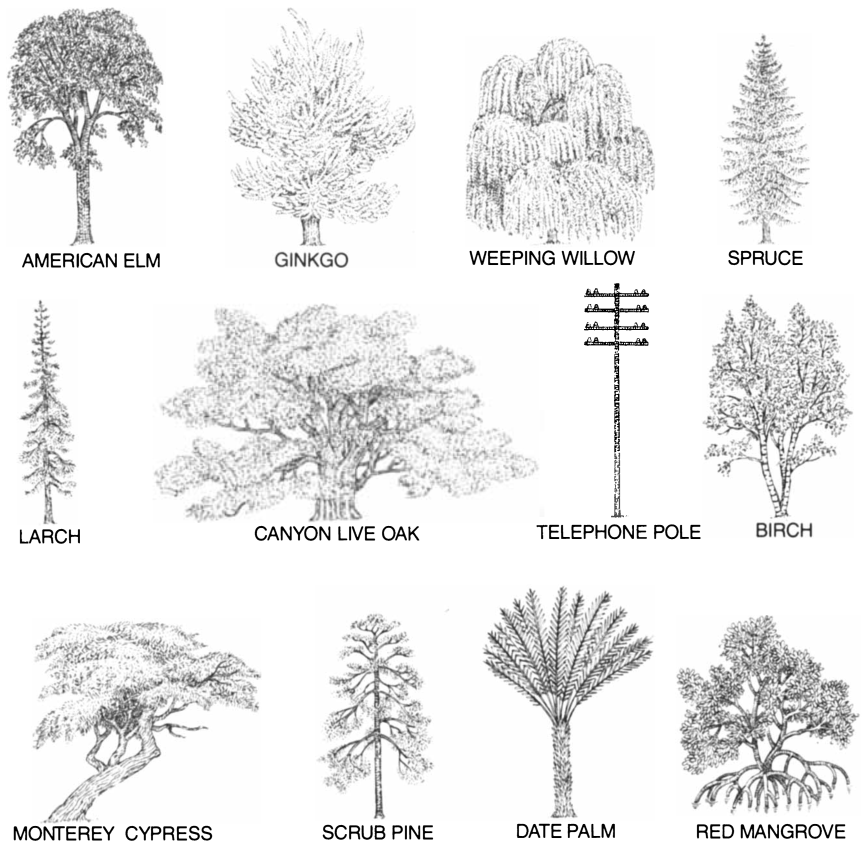
Among the inhabitants of the earth a universal understanding of what is meant by the term "tree" arises from a vast accumulation of common experience. A computer, like an extraterrestrial visitor, cannot draw on this reservoir of common experience; everything must be spelled out for it precisely and unambiguously. Although many properties of trees and other visual scenes have a fair amount of regularity, there is a major component of irregularity that does not fit any simple mathematical or algorithmic model. Any generalized definition that is grounded on an underlying regularity among trees runs the risk of subsuming objects that are not trees. Indeed, the only definition that would not assume any prior knowledge of trees and that would include all trees and exclude all other objects would amount to a description of all types of trees. It is important to recognize that this difficulty is inherent in random problems; it is not just a symptom of fuzzy thinking by human programmers or a poor choice of descriptions.

A simple algorithm will therefore never serve to solve a random problem, because an algorithm for the solution of a random problem would be tantamount to a definition of the problem, and hence it would have to contain all the problem's many possible solutions. For example, an algorithm to identify fingerprints would have to amount to a list of all possible fingerprints, but there is no way to pack such a list in a few lines of computer code. Fingerprints must ultimately be classified into a large number of basically unrelated types—each of which must be considered in order to identify a given print. The solution to random problems therefore lies essentially in memorizing all possible solutions.

Optical technology offers a potentially massive memory, but this alone does not suffice for a practical system to solve random problems. It would be pointless to store the vast data base of a random problem optically, only to search through it sequentially whenever a solution that fits the input data is needed; it would take a prohibitive amount of time. Moreover, the input data as well as the stored information are likely to be incomplete or inaccurate, precluding an exact match between them. The key additional ingredient for a practical system that solves random problems is



OPTICAL MEMORY DEVICES can be made from disks containing embedded spots that modulate light into two possible states. The states correspond to the value of a stored bit of binary data (either a 1 or a 0). In most current designs (*left*) the stored information is accessed serially by focusing a light beam on each data-storing spot and detecting the reflected signal. A similar device (*right*) with an unfocused beam and a transmissive disk could greatly increase the speed with which stored data is accessed: it would scan millions of spots at a time, simultaneously reading out and transferring the data to an array of detector elements.



ARE ALL THESE OBJECTS TREES? Even a young child can answer correctly; a conventional computer, however, has enormous difficulty in doing so. Although there is a fair amount of regularity among the trees shown (each has a trunk and branches, for example), there is also a major component of arboreal irregularity among them. A generalized definition of a tree based on the underlying regularity could lead to erroneous identifications (such as mistaking a telephone pole, which has a "trunk" and "branches," for a tree). Hence any effective program designed to recognize trees would essentially have to be a list of all types of trees, which cannot be done in a few lines of computer code.

a way to associate input data directly with the stored information without requiring an exact match.

Such a process of association is a major feature of biological memory, where partial features of an object trigger the retrieval of complete information about the object. Consider the train of associated reminiscences that courses through one's mind when one sees a familiar face: the person's name, one's general disposition toward him or her and perhaps the smell of his cologne or her perfume—to name a few. Similarly, human beings do not consciously follow an explicit step-by-step algorithm to recognize visual scenes; rather, they follow an unconscious process of association. Even in the case of highly structured problems, such as chess playing, experts develop skills that are associative in nature. (In fact, it is the inability of expert chess players to record explicitly the "algorithm" by which they made a brilliant move that so far has prevented the writing of a chess-playing program capable of beating world-class players.)

Can the anatomical structure of the brain provide an organizational principle by which associations can be readily established between what is stored in memory and the input data? Moreover, can such a model be implemented by taking advantage of the intrinsic strengths of optical technology?

The brain consists of a very large number of neurons each of which is directly connected to a large number of other neurons. A neuron can be in one of two states (known as "firing" or "not firing") and is able to sense the states of its neighbors through its connections. During the course of cerebral "computation" each neuron independently examines the states of its neighbors and, based on the information, determines its own future state. Such a network of neurons is robust; if some neurons malfunction, the overall function of the network is not affected. (Indeed, neurons in the brain are continually dying off, and yet thought and memory are not appreciably hampered.) Computation in neural networks is done in a collective manner: the simple, simultaneous operation of individual neurons results in the sophisticated function of the neural network as a whole.

This form of organization enables thousands of neurons to collectively and simultaneously influence the state of an individual neuron according to the application of simple rules. More important, it also allows information to be encoded in the neural connections rather than in separate memory elements. Each distinct piece of stored information can be represented by a unique pattern of connections among neurons.

Computers whose processing elements are arranged in much the same way as neurons are arranged in the brain would exhibit several features making them remarkably suitable for the solution of random problems. For one thing, such neural computers would be versatile, since the connections between the elements (of which there are a huge number) serve as the programmable storage mechanisms that uniquely "tune" the computer's memory to a given problem. Essentially the connections in a neural computer could be reconfigured in a great many ways to make possible the storage of a random problem's many possible solutions.

Another major feature of the operation of neural computers is spontaneous learning. Imagine what it would be like if children had to be taught how to speak as they are taught how to carry out long division, that is, by teaching them a set of specific rules! Fortunately this is not necessary in most instances, since a child spontaneously associates spoken language with an experience. Learning to talk therefore begins as a process of mimicking the words heard in association with a particular experience. In this simple way the child starts to produce recognizable and sensible patterns of speech.

Similarly, the programmer of a neural computer does not have to understand in a formal, mathematical sense the problem for which he or she is programming. The programmer only has to provide enough "training" data (consisting of possible solutions) to the computer and allow it to set up a unique pattern of connections for each solution. In other words, it is possible for a neural computer to program itself. For example, if one wanted to program a neural computer to recognize different trees, one would provide images of trees as training, allowing a specific pattern of interconnections among the computer's processing elements to be "imprinted" for each training image.

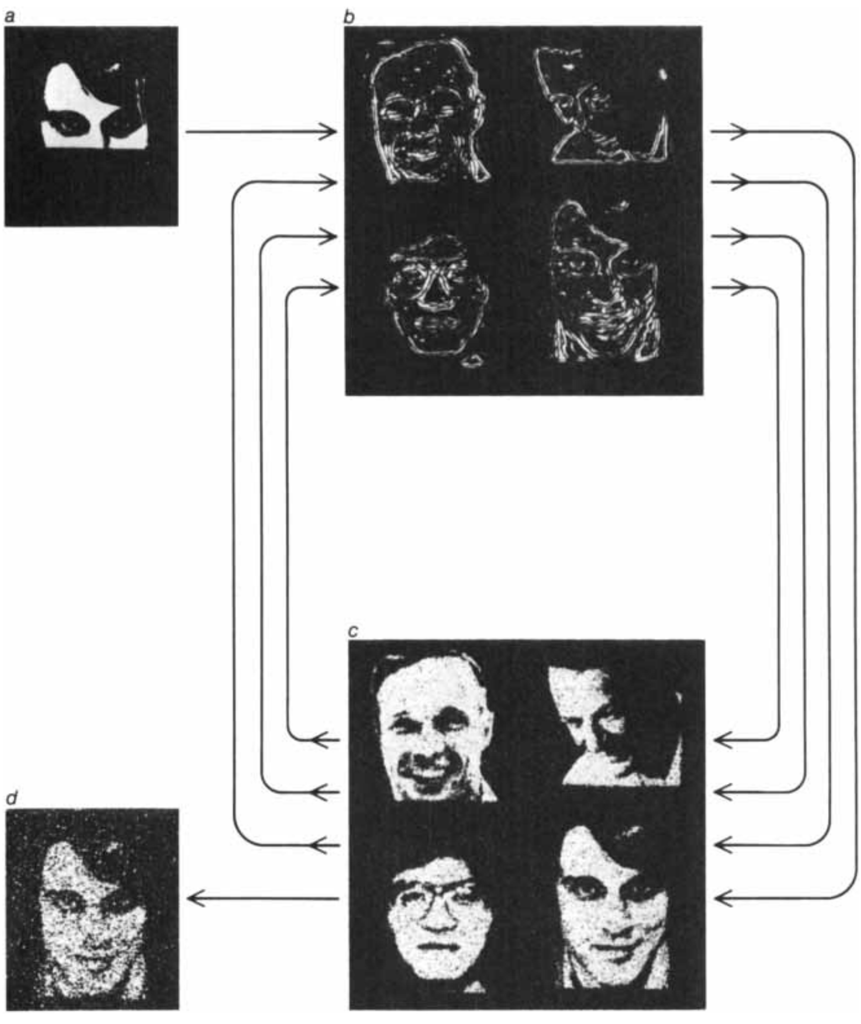
A neural computer built along these lines from optical elements consists of two main components. The first component is a two-dimensional array of optical switching elements to simulate neurons; the elements switch states depending on the states of the elements to which they are connected. Each element in the planar array can be interconnected to all other neurons by light beams. The second component is a hologram that specifies the interconnections among the elements. Since the connections constitute the memory, they must be modifiable—if different problems are to be tackled by a single optical neural computer.

The array of switching elements can be made by well-established fabrication methods developed for semiconducting materials. Each element can be either a purely optical switch or an optoelectronic combination of light detector, electronic switch and light emitter. The total number of possible connections is the square of the number of elements. If a volume hologram is used to specify the interconnection scheme, the volume of the crystal must be proportional to the total number of connections.

A hologram whose volume is one cubic centimeter can in principle specify more than a trillion connections, which means it can handle all possible interconnection schemes of more than a million optical elements. The ability to store the interconnection information in the three dimensions of a volume hologram creates a huge potential memory for optical neural computers. In the case of holograms used for pattern-recognition systems, for example, the interconnection scheme can easily be set up by making a hologram of all the images that are to be identified.

Several experiments are under way in our laboratory at Caltech to develop such optical neural computers. In one experimental setup [*see illustration on page 89*] the action of a two-dimensional array of more than 10,000 neurons is simulated by a threshold device consisting of 10,000 tiny elements that switch the reflectivity of their front surface whenever the intensity of a light beam hitting their back surface is greater than a certain threshold. In this sense the threshold elements act as neurons because they switch states depending on whether enough light reaches them from behind. A pair of planar holograms, a system of lenses and mirrors and an array of pinholes specify how much light each threshold element gets, in essence establishing the interconnections among the elements. Both holograms contain the same set of images, although the edges of the images stored in one hologram have been enhanced. The system is arranged in the form of an optical "loop" so that the system has continuous feedback.

An image to be recognized is projected into the system by reflecting it off the front of the threshold device. Lenses, mirrors and the pinhole array enable the reflected input image to interact with all the images stored in the two holograms in such a way that the best match between the input image and the stored holographic images is the brightest image issuing from the second hologram. The light from the second hologram is then directed to the back surface of the threshold device, causing individual threshold elements to change their reflectivity so that an image of the best match is primarily what is reflected off the front of the device for a second pass around the loop. Successive passes around the loop continue to reinforce the best match until the system "locks" onto the correct stored pattern, which can be retrieved as the output image. In this way the system is capable of recognizing any one of the stored images—even if only part of the image is projected into the system.



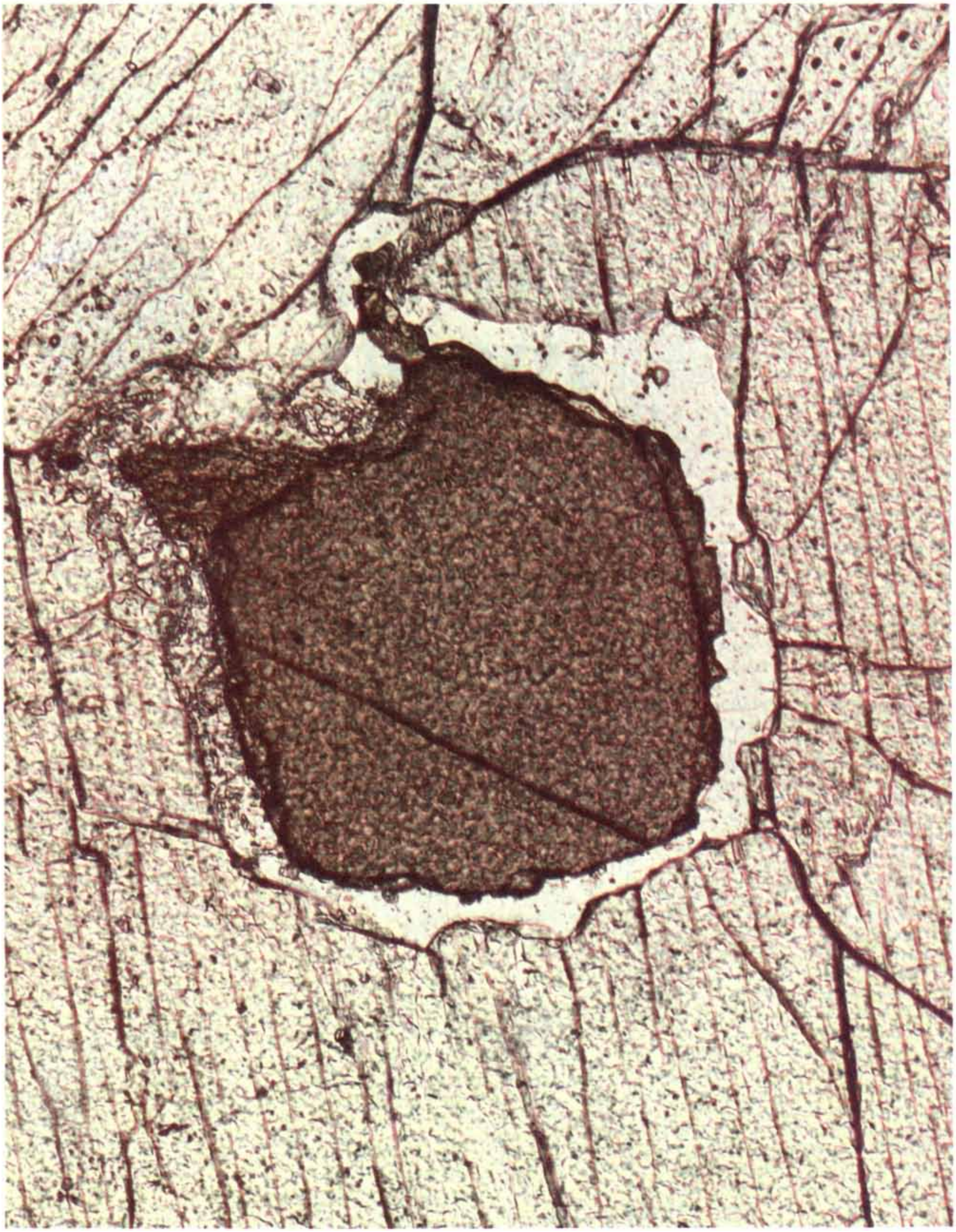
ASSOCIATIVE MEMORY is a feature of the authors' optical pattern-recognition system (*see illustration on page 89*): the system can "recognize" an image even if only part of it is projected into the system. If only half (*a*) of one of the four faces that are respectively stored as edge-enhanced images (*b*) and normal images (*c*) on a pair of holograms is projected into the system's self-reinforcing optical-feedback loop, the system nonetheless selects the correct whole image (*d*) as output. A similar process, in which one piece of information elicits the recollection of related stored information, is a feature of human memory and learning. Although only four images were recorded holographically on film in this example, a volume holograph, which has an enormous memory capacity, could conceivably make possible the memorization and swift recognition of millions of images.

We believe the best way to design computers that solve random problems is through the implementation of a neural architecture. Optical technology can be applied amazingly well in building such a computer. A neural computer demands a very large number of switches, each of which must perform only a simple operation in order to switch between two states. Similarly, it is possible to place a large number of simple optical switching elements on a plane. A neural computer requires extensive connectivity and data communication. Similarly, holograms can establish the necessary connections between numerous optical elements. Because light rays can cross one another without interfering and are not limited to traveling along the two dimensions of the surface of a silicon chip, simultaneous communication among numerous optical elements is readily achieved.

Although optical systems are vulnerable to various inaccuracies and local errors, neural computers are inherently fault-tolerant: a perfect match between input and output is not necessary. A neural computer would be programmed by establishing a unique interconnection pattern for each solution, and this could easily be done by recording various training images on a photorefractive crystal or photographic hologram. In contrast to conventional computers, the speed of the individual switching elements is not critical for the function of a neural computer, since a few iterations usually suffice to complete the association function. This is particularly fortunate for optical neural computers, since each firing of a "neuron" consumes a fixed amount of energy and speeding up translates directly into more power consumption and hence into excessive heat production.

Clearly many challenges must be met before optical hardware, arranged in a neural architecture, can produce practical computers that are capable of dealing with random problems. Advances in optical materials and manufacturing technologies and in the understanding of the organization of large-scale neural computers are needed. Equally important is better understanding of the operation of neurons in the brain and of how they collectively "learn" and "classify" patterns.

Engineers, computer scientists and mathematicians have reached significant turning points in three seemingly unrelated areas: optical components, neural computers and random problems. There is good reason to believe that, with progress in each of these areas, their interaction will ultimately yield systems capable of pattern recognition and other artificial-intelligence tasks that may never be duplicated by purely electronic means.



RIFT EVOLUTION is reflected in the chemical composition of rocks from the earth's mantle that have been exposed at the surface. The micrograph shows a single grain of spinel (*dark brown*), a minor mineral component of peridotite, which is the dominant rock in the mantle. The grain is about half a millimeter in diameter; it is embedded in pyroxene, one of the chief minerals in peridotite. The ratio of chromium to aluminum in spinel increases as a rift evolves from the continental toward the oceanic stage. This suggests that the degree of melting of mantle peridotites under the rift increases with time; melting tends to extract more aluminum than chromium from spinel. The cause of the increase is probably an intensification of an underlying thermal anomaly in the mantle.

The Rifting of Continents

It begins above a hot zone in the mantle. Upwelling molten rock "underplates" and weakens the continental crust, piercing it at discrete points and finally rifting it in two: an ocean is born

by Enrico Bonatti

Since the theory of continental drift gained wide acceptance about 20 years ago, investigators have been preoccupied by the following question: How does a continent split open, allowing an ocean to form in the gradually widening rift? The problem is more complex than it might first appear. The lithosphere—the earth's rigid outer layer, including the crust and part of the upper mantle—is very different under oceans from what it is under continents. Whereas continental lithosphere is generally between 100 and 150 kilometers thick, the thickness of oceanic lithosphere increases with its age, ranging from less than 10 to no more than 100 kilometers. Continental crust in particular is thicker than oceanic crust, and it is also less dense: it consists primarily of granitic rocks, which are rich in silicon and aluminum and have a relatively low density of about 2.7 grams per cubic centimeter. Oceanic crust, in contrast, consists primarily of iron- and magnesium-rich rocks of basaltic composition, whose density is about 2.9 grams per cubic centimeter. When continental lithosphere rifts into two separate plates, thick lithosphere must somehow give way to thin lithosphere; granite must give way to basalt. What physical and chemical processes are involved?

Nowhere can these problems be studied better than in the region of East Africa and Arabia, where some of the earth's major rifts are found. The great valleys that extend from Mozambique and Zambia north to Ethiopia make up what is called a continental rift system, in which the rupture of the continental lithosphere is not yet complete. The Gulf of Aden, on the other hand, is an oceanic rift: there the ruptured continental blocks, Arabia and Africa, have been diverging for more than 10 million years, during which time hot magma from the underlying mantle has been welling up between the blocks to form oceanic crust. In the Red Sea the transition from a continental rift to an oceanic rift is taking place right now.

It seems safe to assume that these three rifts represent different stages of a single evolutionary sequence. In other words, the Gulf of Aden evolved from a rift of the Red Sea type, which in turn developed from a continental rift system like the one in East Africa. A more advanced stage of this hypothetical sequence would be the Atlantic Ocean, a mature rift that over more than 100 million years has evolved from conditions similar to those that prevail today in the Red Sea and the Gulf of Aden. Thus the young rifts of East Africa and Arabia have a particular geologic significance: by studying them one can hope to learn how Europe and Africa initially separated from the Americas.

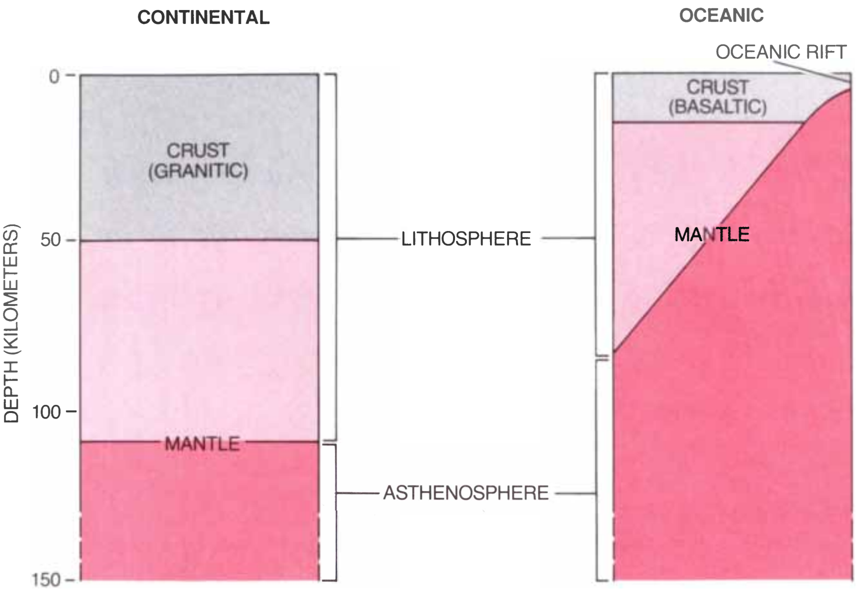
One thing is certain about the rifting process: the upper mantle is intimately involved. Seismic studies have revealed that under the lithosphere there is a zone of the mantle in which acoustic waves (triggered by earthquakes or by manmade explosions) travel slower than they do in the lithosphere. The low-velocity zone, called the asthenosphere, must be made of hotter material than the lithosphere; the hotter material is weaker and less rigid and therefore does not transmit acoustic waves as well. The seismic investigations have shown that under a major continental rift, such as the East African rift, the Rio Grande rift in the western U.S. or the Baikal rift in the Soviet Union, the thermal boundary between the lithosphere and the asthenosphere rises sharply; that is, the lithosphere gets thinner and the asthenosphere wells up. Under the East African rift, for example, the boundary lies at a depth of only between 30 and 50 kilometers, whereas elsewhere in Africa the lithosphere is between 100 and 150 kilometers thick.

There are two schools of thought on what this means, embodying different views of what causes rifting. Some workers believe rifting begins in the lithosphere, when stresses arising from the horizontal motions of the lithospheric plates stretch one of the plates to the breaking point. In this view the asthenosphere is essentially passive: it simply rises to fill the gap left by the thinning lithosphere.

In contrast, the second school holds that the rifting process begins in the asthenosphere with the establishment of a thermal anomaly—a hot spot or line—in which the temperature is higher than it is at an equal depth elsewhere [see "The Earth's Hot Spots," by Gregory E. Vink, W. Jason Morgan and Peter R. Vogt; SCIENTIFIC AMERICAN, April, 1985]. Because it is anomalously hot, the mantle rock under the nascent rift wells up, swelling and thereby stretching the overlying continental lithosphere. At the same time the strength of the lithosphere is reduced by the heat. Eventually the upper part of the weakened lithosphere cracks along extensional faults, and blocks of crust drop down along them. The result is a series of grabens, or rift valleys, like the ones in East Africa.

To a certain extent the question of which comes first—the stretching of the lithosphere or the upwelling of the asthenosphere—is a chicken-and-egg question. On the geologic time scale both processes occur more or less simultaneously. Nevertheless, the view that emphasizes mantle upwelling as the ultimate cause of rifting is supported by research I am currently carrying out with my colleague Monique Seyler at Columbia University's Lamont-Doherty Geological Observatory.

Our research has focused on determining what happens to the upwelling mantle rock under a rift. Since pressure within the earth increases with depth, the hot material under a



CONTINENTAL LITHOSPHERE AND OCEANIC LITHOSPHERE differ in thickness and composition. The lithosphere includes the crust and part of the upper mantle. It is divided into rigid plates that move over the underlying asthenosphere, a layer of the mantle that is hotter and weaker than the lithosphere. Continental lithosphere is between 100 and 150 kilometers thick; in contrast, oceanic lithosphere is never more than 100 kilometers thick, and at oceanic rifts, where it is created by the upwelling of the asthenosphere, its thickness is less than 10 kilometers. Continental crust is mostly granitic; oceanic crust consists of basalts and gabbros, both of which form from basaltic magma.

rift will be subject to lower pressures as it rises. Such a pressure drop, with no significant drop in temperature, is likely to cause a fraction of the material to melt. The molten rock, or magma, separates from the unmelted fraction of the parent rock and continues to rise. Ultimately it does one of two things: either it erupts volcanically at the surface or it cools and solidifies slowly in magma chambers within or below the crust, forming intrusions of igneous rock.

Evidence that the latter process has taken place under the Red Sea rift can be found on the island of Zabargad. The island consists entirely of blocks of upper mantle and lower crust that during the rifting process have somehow (exactly how is not clear) been lifted up and exposed at the surface. Among the exposed blocks are peridotites: dense (3.3 grams per cubic centimeter) rocks that make up most of the upper mantle and whose chief mineral components are iron-magnesium silicates such as olivine and pyroxene. In contact with the peridotite is a section of granulitic gneiss, a granitic rock characteristic of the lower part of the continental crust. The association of mantle-derived peridotite and granulitic gneiss suggests that Zabargad is a rare resource indeed, a place where one can actually look at pieces of the crust-mantle boundary.

A close look reveals important evidence about how rifting began in the Red Sea. The gneisses on Zabargad are intermingled with gabbros: coarse-grained igneous rocks formed from basaltic magmas that have cooled slowly at depth. (When the same magmas erupt onto the surface and cool rapidly, they produce fine-grained basalts.) Laboratory experiments have demonstrated that basaltic magmas—and hence gabbros and basalts—are derived from the partial melting of peridotite in the upper mantle.

By examining the mineralogy and chemistry of a gabbro, in particular by analyzing the elemental composition of its constituent pyroxenes, one can infer the depth at which it cooled and crystallized. The Zabargad gabbros crystallized at a depth of at least 30 kilometers below the surface of the crust. The intermingled gneisses must have come from about the same depth, because they include garnet, a mineral that crystallizes only under high pressure. Since the continental crust in the vicinity of the Red Sea is between 30 and 45 kilometers thick, the Zabargad gneiss-gabbro complex must have been created by the injection of basaltic magmas at the base of the crust before any significant rifting or thinning of the crust had taken place.

That means the upwelling of the asthenosphere must have preceded the thinning of the crust. Indeed, the gneisses and gabbros bear witness to the thinning process. Their mineral structure and chemistry indicate that after they crystallized they underwent metamorphic change and recrystallization under conditions of gradually decreasing pressure. In other words, they were gradually lifted to shallower depths as the continental crust was stretched and thinned.

On other islands in the Red Sea my colleagues and I have found gabbros that initially crystallized at depths of less than 10 kilometers. Apparently basaltic melts from the mantle were emplaced at the bottom of the crust throughout the period during which the crust was being stretched and thinned. As a result a more or less continuous layer of gabbros "underplates" the continental crust of the Red Sea, from the thick, nearly unstretched crust at its shores to the thin crust near the rift axis. Underplating is taking place now under the East African rift; several workers, notably R. C. Searle of the British Institute of Oceanographic Sciences, have reported seismological evidence for the existence of a gabbro layer (gabbros have a characteristic seismic velocity) under the crust in the rift valleys.

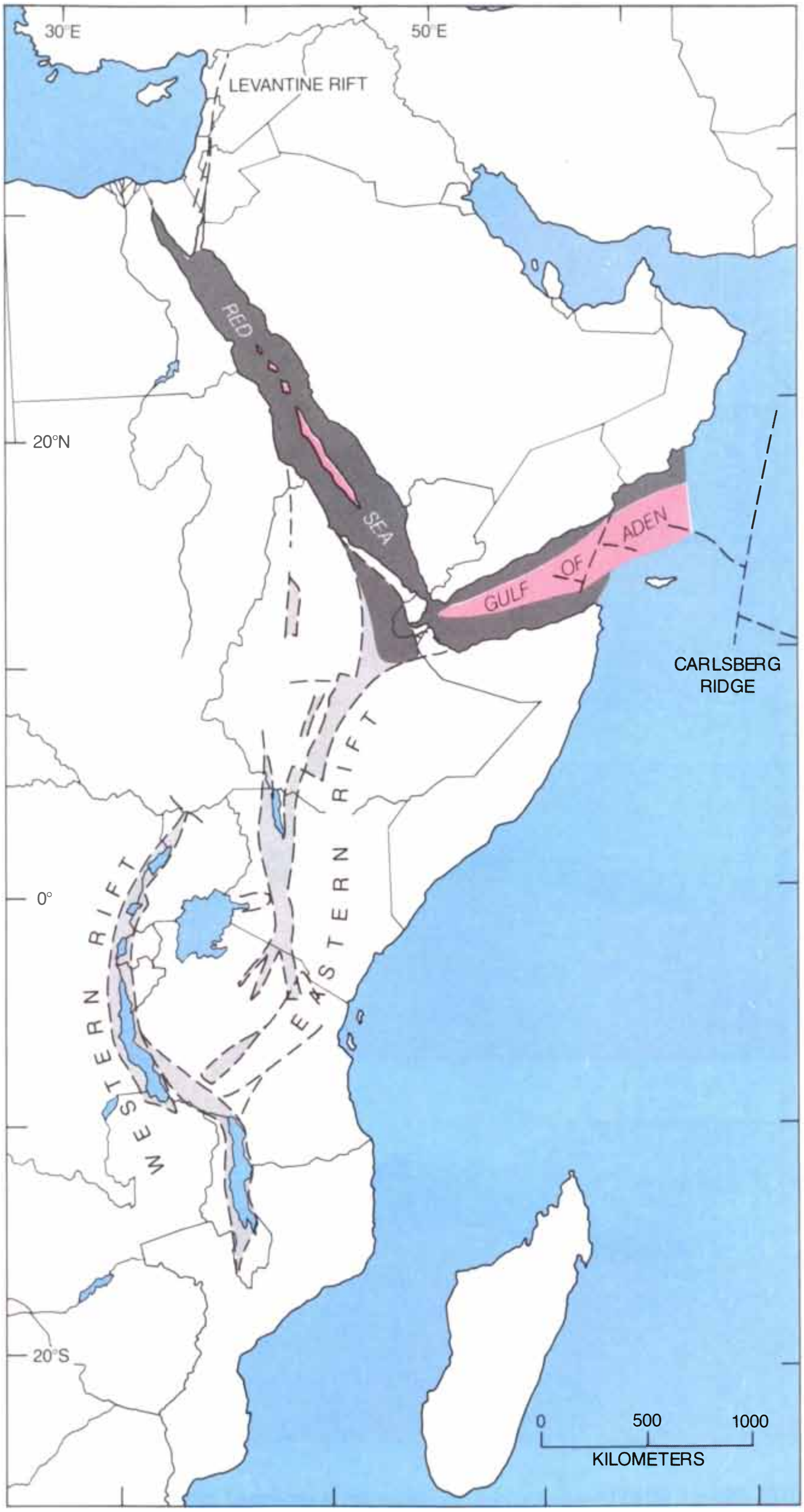
It is remarkable that the formation of the Atlantic rift also seems to have begun with underplating. Seismic profiles of the crust off the east coast of North America (made by a multi-institutional group of scientists from the U.S. and Canada as part of the Long-Aperture Seismic Experiment) indicate that a layer of gabbroic rocks underlies the entire continental shelf. Toward land the gabbro layer dips downward and probably reaches under the thick continental crust, as the analogous layer does under the Red Sea. Away from the shore, where the continental crust thins and is finally replaced by oceanic crust, the gabbro layer rises to progressively shallower depths. Hence the breakup of Europe and North America, which took place roughly between 170 and 100 million years ago, appears to have been accompanied by the same upwelling of basaltic magma now going on under the East African and Red Sea rifts.

The underplating of the continental crust by basaltic magma heats and weakens the crust. A fraction of the magma penetrates faults and fissures in the crust, and occasionally some of it erupts at the surface. As the crust gets thinner eruptions become more frequent; moreover, the injections of magma tend to become concentrated in a narrow zone around the rift axis. Finally the width of the zone shrinks to just a few kilometers, within which the last remnants of continental crust have been swept away by upwelling mantle rock. The continent has split, and oceanic crust, consisting simply of cooled basaltic magma, is filling the crack.

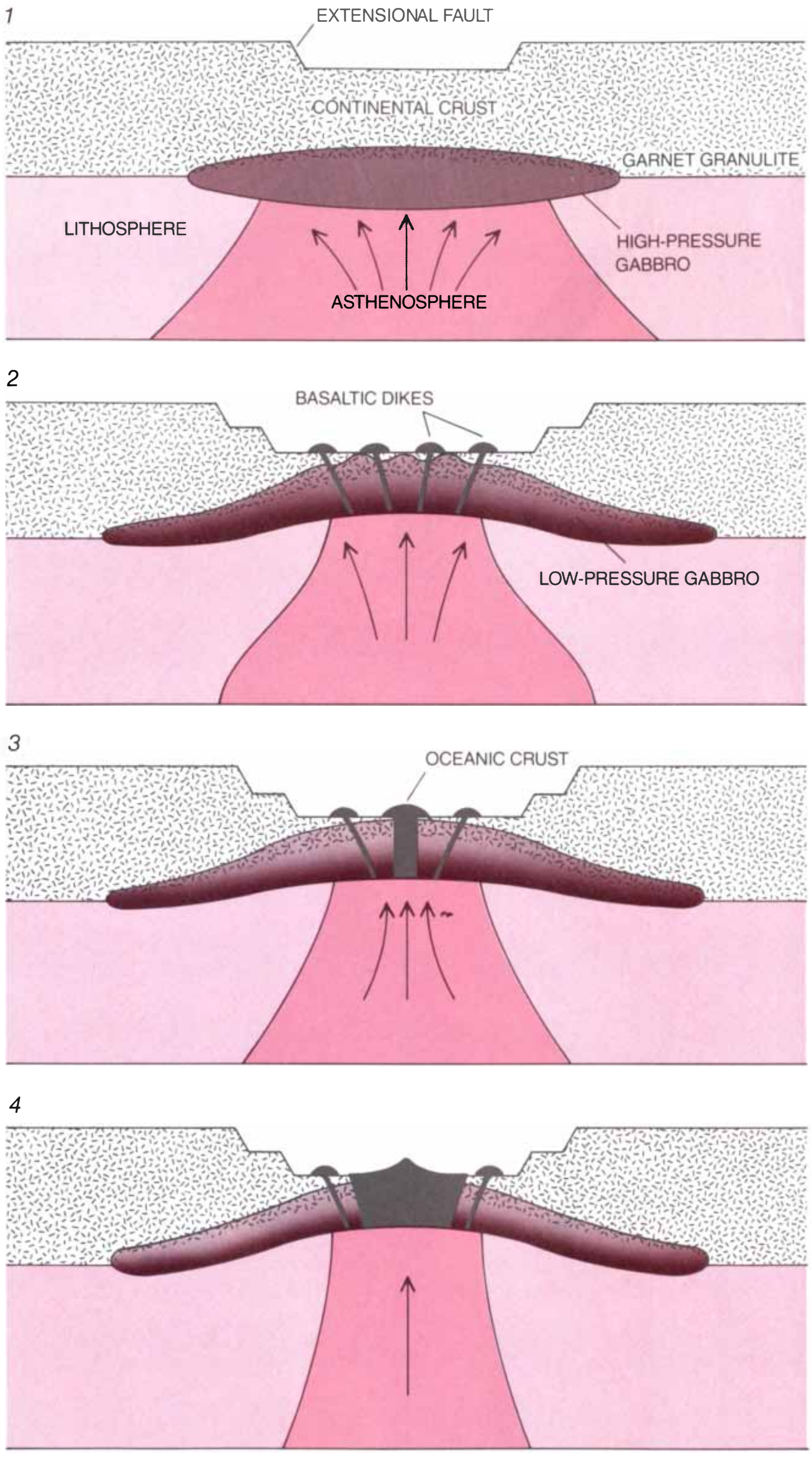
The top layer of the oceanic crust is made of basalts formed by magma that has erupted onto the surface. Below the basalts and extending to the base of the crust are gabbros that have cooled slowly in magma chambers under the rift. Below the gabbros lie the peridotites of the mantle. The peridotites under the rift differ from peridotites elsewhere in the mantle in that they have been depleted of the melt fraction that formed the basalts and the gabbros. Initially the depleted peridotites belong to the asthenosphere because they are hot, but as the two newly separated plates spread away from the rift, the asthenospheric material adheres to the edges of the plates, cools and is thereby transformed into new oceanic lithosphere. (That is why oceanic lithosphere gets thicker as it gets older.)

Oceanic crust can be distinguished from continental crust on a broad scale, without dredging up a lot of rock samples, on the basis of its magnetic properties. (The fact that crust is under water does not prove that it is oceanic; continental shelves, the stretched edges of continents, are also under water.) When magma erupting from an oceanic rift cools to form basalt, the magnetic minerals that crystallize from the magma align themselves with the earth's magnetic field. They remain locked in that orientation when the crust later spreads away from the rift and even when the magnetic field switches polarity, as it does at irregular intervals of a few hundred thousand years or so. The result is a series of magnetic anomalies arranged in stripes parallel to the rift axis. The striped anomalies, which can be measured with instruments towed by a research vessel, are a signature of oceanic crust.

Data from such magnetic surveys suggest that the transition from a continental to an oceanic rift is now under way in the central Red Sea. In the northern Red Sea there are no clear-cut striped magnetic anomalies, indicating that the continental crust has not yet been rifted; to the south there are linear anomalies associated with a trough that runs along the axis of the sea, indicating that the sea floor there has been spreading for a few million years. In the central Red Sea, on the other hand, there are strong linear magnetic anomalies, but only in a few isolated areas along the axis. The anomalies are associated with seafloor trough segments that are spaced at regular intervals of about 50 kilometers. As one moves from south to north the magnetic anomalies become



RIFT SYSTEMS of East Africa and Arabia exemplify different stages in the evolution of a rift. The valleys of East Africa (*light gray*) form a continental rift system: the lithosphere there has been thinned, allowing blocks of crust to drop down along extensional faults, but it has not yet ruptured. In the Gulf of Aden, an oceanic rift, the continental lithosphere ruptured some 10 million years ago, and since then basaltic magma has been welling up from the asthenosphere to form oceanic crust (*red*). The transition to an oceanic rift took place more recently in the southern Red Sea, and it is going on now in the central Red Sea, where oceanic crust is forming at isolated points. Most of the Red Sea floor (and part of the African coast) consists of stretched continental crust (*dark gray*).



STAGES OF RIFTING posited by the author are diagrammed in vertical cross section. The first stage in the evolutionary sequence is typified by the East African rift valleys (*1*), where the continental crust is being underplated by gabbros. The gabbros are crystallizing under high pressure and are mingling with the garnet granulites that make up the base of the continental crust. In the northern Red Sea (*2*) the crust has already been stretched and thinned considerably. The gabbros at the base of the thinned crust are chemically distinct because they have crystallized at shallower depths and thus under lower pressures. Basaltic dikes have also been injected into fissures in the crust. In the southern Red Sea (*3*) the zone of melt injection has become concentrated at the rift axis; the African and Arabian continental blocks have split, and oceanic crust is forming in the crack. In the Gulf of Aden (*4*) the sea floor has been spreading for 10 million years.

more subdued, and the corresponding trough segments become shorter and narrower.

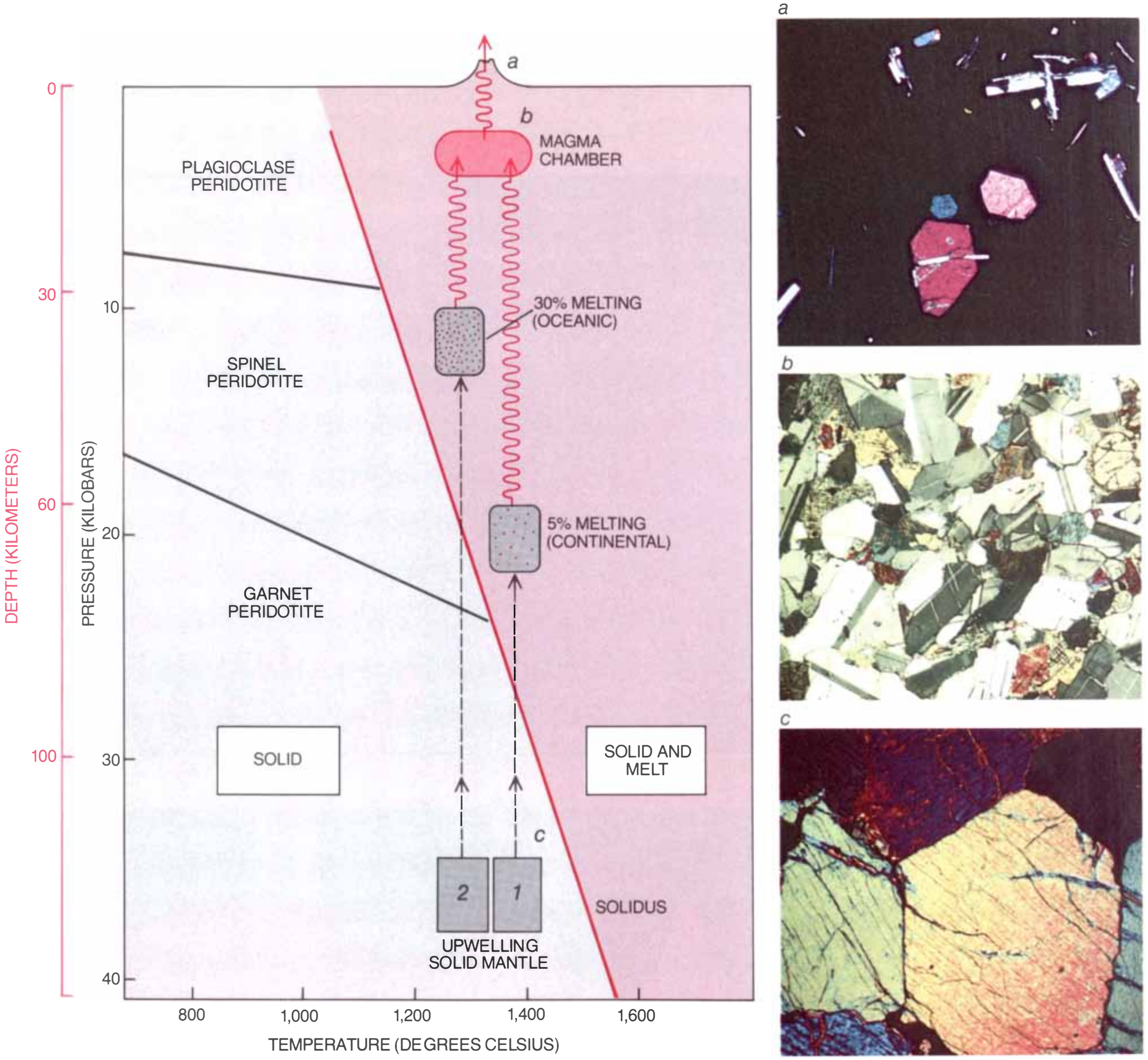
The observations suggest two conclusions. First, the rifting of the Red Sea is proceeding from south to north. Second, in the central Red Sea the transition from a continental to an oceanic rift and the initial emplacement of oceanic crust have taken place not along a continuous axial crack but at discrete, regularly spaced points. Subsequently the zones of crust formation have grown from points into linear segments, producing the observed troughs on the sea floor. Eventually the linear segments will merge into a long axial zone of sea-floor spreading like the one already present in the southern Red Sea.

If that model is correct, one would expect to find magnetic and topographic discontinuities at regular intervals along a spreading axis; the initial spreading nuclei are not necessarily arranged in a straight line, and so the axis may be offset slightly at the junctions between different linear segments. Searle and Z. Garfunkel and A. Ginzburg of the Hebrew University of Jerusalem have found such discontinuities, spaced roughly 50 kilometers apart, in the southern Red Sea. Moreover, Hans Schouten and Kim D. Klitgord of the Woods Hole Oceanographic Institution have made similar discoveries in the Atlantic. Specifically, they have detected equidistant breaks in the magnetic-anomalies pattern in an area of the western Atlantic that formed between 155 and 108 million years ago, during the earliest stages of sea-floor spreading. The beginning of sea-floor spreading in the Atlantic, like the initial rifting of Europe and North America, seems to have been driven by processes similar to those now at work in the Red Sea.

Why does sea-floor spreading begin at discrete points? The answer must have something to do with the pattern of upwelling and partial melting of the asthenosphere under a rift. Because the rising asthenosphere is partially molten, and because it may contain volatile compounds such as water and carbon dioxide, it is less viscous and less dense than the overlying lithosphere. This situation is gravitationally unstable: low-density material tends to float on high-density material rather than vice versa. The instability can be modeled experimentally with two liquids of differing densities and viscosities. Bruce D. Marsh of Johns Hopkins University and John A. Whitehead, Jr., of Woods Hole have done such experiments, and the results are rather striking. When the liquid of lower density and viscosity is injected under the other liquid, the initially horizontal boundary between the two substances almost immediately develops undulations. Soon thin plumes of the lower liquid begin rising into the upper liquid at regularly spaced points. The plume spacing, it turns out, is proportional to both the thickness of the lower liquid layer and the ratio of the viscosity of the upper layer to that of the lower layer.

Extrapolating from the laboratory models to the earth's rift zones is obviously risky. But the experimental results do suggest that the reason sea-floor spreading begins at discrete points is that the upwelling of partially molten asthenosphere under a rift takes the form of thin diapirs, or plumes. On the basis of the experiments one can further predict that the spacing of the diapirs should depend on the thickness of the melt-containing layer: the thicker the layer, the wider the spacing. In addition the spacing should be proportional to the degree of melting within the melt layer, since the higher the melt content, the lower the viscosity of the layer.

What this means is that the spacing between diapirs should grow with the intensity of the thermal anomaly—which feeds many diapirs—deep in the mantle under a rift. A number of observations indicate that the spacing of zones of volcanic activity tends to increase from about 50 kilometers along a continental rift to as much as 150 kilometers along an oceanic one. Furthermore, the spacing along an oceanic rift is proportional to the sea-floor

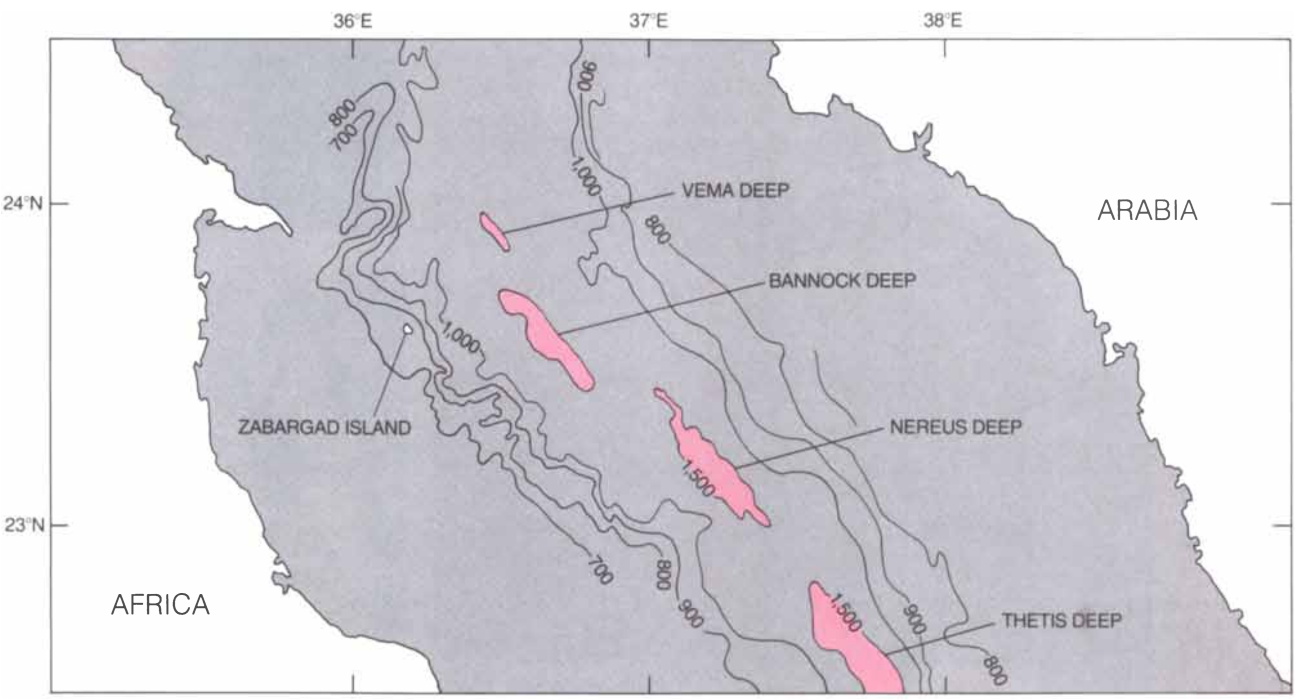


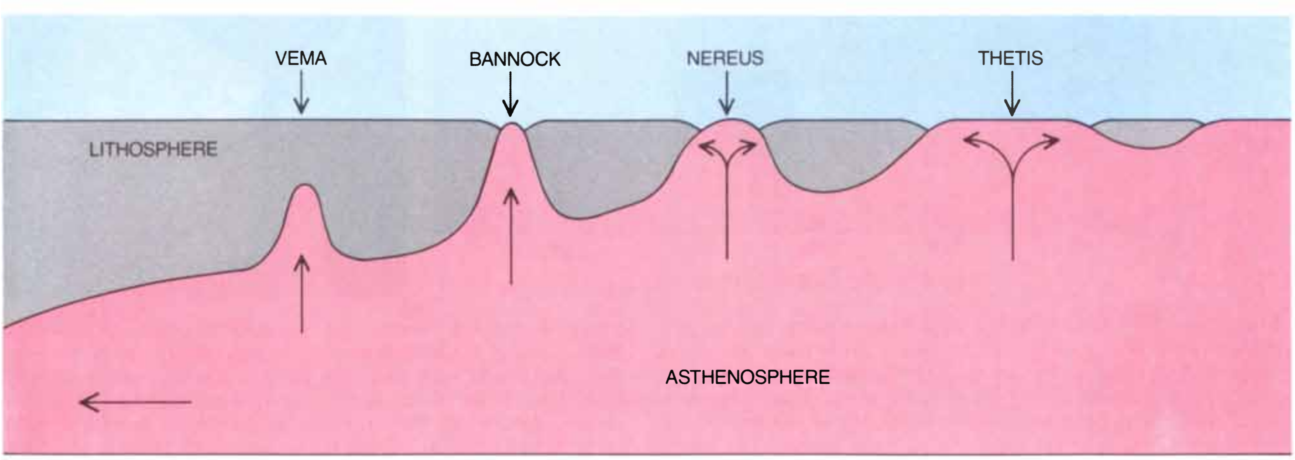
PARTIAL MELTING of mantle rock under a rift occurs because the upwelling rock is subjected to progressively lower pressures, while its temperature remains roughly constant until it nears the surface. The diagonal line on the temperature-pressure diagram is the "solidus" line for peridotite (*c*): to the left of the line peridotite is totally solid, to the right it is partially molten. The degree of melting increases as pressure decreases. Under a nascent continental rift (*1*) the partially molten peridotite rises to the base of the continental crust—to a depth of about 60 kilometers—and roughly 5 percent of the rock melts. Under an oceanic rift (*2*), where the crust is thinner, the hot peridotite rises to shallower depths, and as much as 30 percent of it melts. The melt separates from the parent rock and rises farther, leaving behind depleted peridotite. Some of the melt cools slowly in magma chambers, forming gabbro (*b*); the rest erupts at the surface and cools rapidly, forming basalt (*a*). Hence the oceanic crust consists of a layer of basalt overlying a layer of gabbro. The composition of both the crustal rocks and the depleted mantle peridotite depends on the degree of melting. For example, basalts found at continental rifts are richer in alkali metals than basalts dredged from oceanic rifts.

spreading rate; for example, surveys carried out by Kathleen Crane of Lamont-Doherty and by Schouten and his co-workers have shown that zones of enhanced igneous injection are considerably farther apart along the East Pacific Rise, where two lithospheric plates are separating at a rate of between 12 and 18 centimeters per year, than they are along the Mid-Atlantic Ridge, where the spreading rate is only two to four centimeters per year.

The field surveys suggest that as a rift evolves from a continental into an oceanic rift, or from a slow-spreading oceanic rift into a fast-spreading one, magma injection becomes more intense but at the same time more concentrated, driven by fewer but larger diapirs. (To be sure, lava erupts along the entire length of an oceanic rift, because the material in the upwelling diapirs migrates along the rift axis below the surface.) On the basis of the laboratory experiments one can therefore conclude that the thickness of the melt layer and the degree of melting in the asthenosphere must increase during the evolution of a rift. The cause of the increase is probably an intensification of the large thermal anomaly in the underlying mantle.

The hypothesis is supported by the chemistry of crustal basalts and mantle peridotites found at rift zones. Basalts from volcanoes in East Africa and at other continental rifts are relatively rich in alkali metals such as sodium and potassium; these elements are among the first to separate from the mantle rock when upwelling begins, because they are readily extracted even by low degrees of melting. As upwelling intensifies and the rift evolves toward the oceanic phase, the melting of the parent peridotite occurs at lower pressures (shallower depths), and the degree of melting increases to as much as 30 percent. The concentration of alkali metals in the magma is diluted by other, less readily extracted elements. Consequently the basalts found at oceanic rifts are relatively alkali-poor.





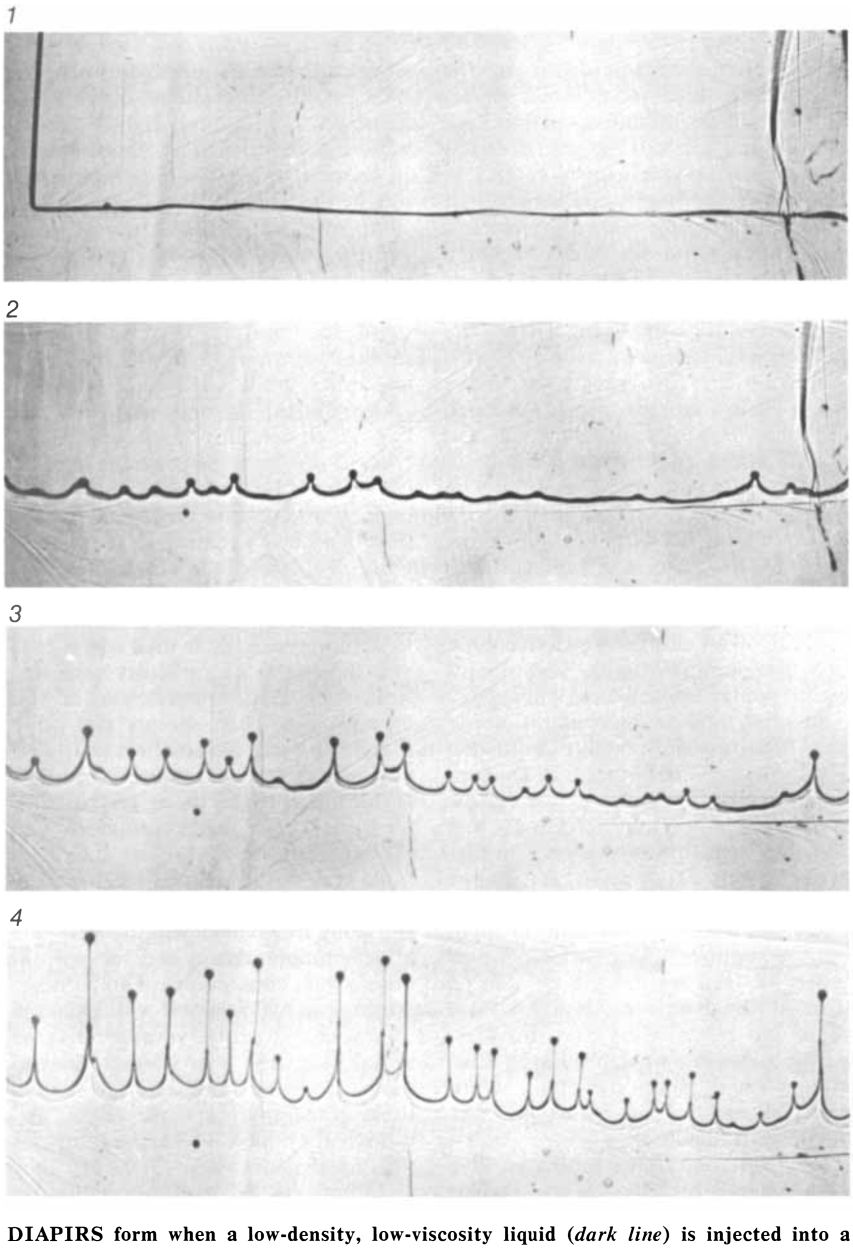
CENTRAL RED SEA depths (in meters) suggest that sea-floor spreading begins at discrete points (*top*). The four deep troughs are associated with striped magnetic anomalies of the type that characterize oceanic crust. Outside the troughs the sea floor consists of stretched and thinned continental crust. The troughs may be underlain by upwelling diapirs in the asthenosphere (*bottom*).

The chemistry of mantle peridotites has been harder to study than that of basalts. One reason is that peridotites are less commonly exposed on the sea floor or at the surface; moreover, the composition of the exposed samples has often been modified by processes taking place near the surface. Since peridotites are the parent rocks of basalts, however, one can say that their evolution during rifting must be complementary. If the degree of melting is higher under oceanic rifts, the mantle must be more depleted of the basaltic-melt fraction there than it is under continental rifts. The difference should be reflected in the mineralogy and chemistry of the peridotites. From laboratory experiments it is known, for example, that peridotites become progressively depleted of the mineral clinopyroxene as they are subjected to higher degrees of partial melting; clinopyroxene, which contains most of the alkali elements in peridotite, breaks down before the rock's other principal minerals do.

In order to estimate the degree of partial melting undergone by a peridotite one must look not only at its mineral composition but also at the chemical composition of the minerals. For instance, the ratio of magnesium to iron in olivine and pyroxene, the chief mineral components of peridotite, increases with the degree of melting. Another sensitive indicator is the composition of the rare mineral spinel, an oxide that contains aluminum and chromium in addition to iron and magnesium. Because aluminum is easily extracted by melting, whereas chromium remains in the solid, the ratio of chromium to aluminum in spinel increases during partial melting. Our measurements show that peridotites from Zabargad have a very low chromium-aluminum ratio, indicating they are relatively unmelted and undepleted, as one would expect at a preoceanic rift. In contrast, mantle peridotites dredged from the Mid-Atlantic Ridge have a relatively high ratio, indicating they have undergone a high degree of melting.

Peridotites from the North Atlantic also display interesting regional variations in the degree to which they have been melted, as Henry J. B. Dick and his co-workers at Woods Hole and my colleague Peter J. Michael and I have shown. In particular, the peridotites from a broad region around the Azores are totally depleted of a basaltic fraction. Not surprisingly, the oceanic crust in that region is unusually thick; it consists, according to Jean-Guy E. Schilling of the University of Rhode Island and E. Klein and Charles Langmuir of Lamont-Doherty, of basalts whose composition suggests they were extracted from the mantle by a high degree of partial melting. Elsewhere in the Atlantic the peridotites are less depleted and the crust is thinner. The Azores, it seems clear, are underlain by an intense hot spot in the mantle.

The evidence I have presented here points to a simple conclusion: the engine driving the birth and evolution of rifts, and consequently the breakup of continents and the formation of ocean basins, are thermal anomalies in the upper mantle. The anomalies are not permanent features. Over geologic time the zones of volcanic activity along a rift are constantly shifting, and so it is perhaps appropriate to think of the underlying anomalies as thermal pulsations or waves. To say that thermal pulsations in the mantle cause rifting, however, is to answer only half of the question. What causes the thermal pulsations? That puzzle has stirred the imagination of many geophysicists. It will continue to do so for some time, because a simple and satisfactory answer has yet to be found.



DIAPIRS form when a low-density, low-viscosity liquid (*dark line*) is injected into a liquid of relatively high density and viscosity. In the experiment illustrated here John A. Whitehead, Jr., of the Woods Hole Oceanographic Institution injected a mixture of water and glycerine into a bath of pure glycerine. Within 45 seconds the water-glycerine mixture began rising in regularly spaced diapirs. A similar phenomenon may explain why sea-floor spreading begins at regularly spaced points: the partially molten asthenosphere under a rift is less dense and less viscous than the lithosphere, and so it rises in diapirs.

Thermoregulation in Winter Moths

Curiously lacking in highly specialized adaptations for the cold, certain nondescript moth species can nevertheless do what their relatives cannot: fly, feed and mate at near-freezing temperatures

by Bernd Heinrich

Winter means death to the adults of many insect species in the middle and upper latitudes of the north. If the cold does not kill them, food shortages often do. Ironically, the same challenges that doom most insects make the northern landscape a refuge for others. Winter's trials force birds and bats, the animals' major predators, to fly south or (in the case of certain bats) to hibernate. The cold even kills some parasites.

Among the few insects that have managed to adapt to a winter existence are some 50 species within the Cuculiinae subfamily of the widespread Noctuidae, or "owlet moth," family. The cuculiinids are dull-colored night fliers populating northern deciduous forests. Their winter-adapted varieties have somehow reversed the typical life cycle of the noctuids.

Most owlet moths are active only on warm summer nights. They die off as the winter approaches, leaving behind eggs, larvae (caterpillars) or pupae that remain inactive until the spring. The winter noctuids, in contrast, emerge as adults in the fall or late winter, when they feed, mate and lay their eggs before dying in the spring. Their caterpillars feed in the early spring (eating the buds of forest trees) and then are quiescent throughout the summer. (Adult winter moths generally feed on the sap of injured trees, although on late-fall nights a few years ago I saw many of them feast on the blossoms of witch hazel, Vermont's latest-blooming plant. Until that time no one knew just how the plant was pollinated.)

How do the winter moths survive when other moths die? What enables them to avoid freezing as they rest, and what makes it possible for them to fly—and so to seek food and mates—in the cold? I tackled the latter question first because I was intrigued by an apparent contradiction: I was sure these moths, like other moths that fly, were endothermic, or able to produce heat by their own metabolism. I also assumed they required high body temperatures in order to fly. Nevertheless, it seemed unlikely that endothermy alone could enable the animals to attain the elevated temperatures needed for flight on cold and sometimes snowy nights.

I suspected the animals had a high flight temperature because summer moths and tropical moths that have a similar body size and wingbeat frequency can fly only when the thorax, which contains the flight muscles, reaches at least 30 degrees Celsius (86 degrees Fahrenheit). On the other hand, different data suggested that the tiny insects (weighing less than .2 gram) would cool too quickly to maintain such temperatures. Small animals, which have a high ratio of surface area to volume, cool faster than larger ones and have greater difficulty retaining heat. Indeed, for many years it was thought that bats, shrews and hummingbirds were the smallest endothermic animals. The minutest of these vertebrates weigh three grams; they are still virtual giants compared with the cuculiinids.

To resolve the matter I first had to capture winter moths, which I did by smearing tree trunks with bait: diluted honey, maple syrup, beer or various other sweet concoctions. Temperature measurements obtained with hair-thin thermocouple probes revealed that my initial suspicion was correct: the insects do require—and generate—a high thoracic temperature for flight. Although they have the same temperature as the air when they rest, they endothermically heat themselves to 30 degrees C. or more before flying, even in air temperatures near zero (the freezing point of water).

Having established that the moths generate their own heat, I attempted to determine whether they have any special physiological tricks for doing so. It was obvious from their vibrating wings that the insects obtain heat by shivering, but this was not in itself special. Ann E. Kammer, now at Arizona State University at Tempe, had shown earlier that Lepidoptera (moths and butterflies) shiver during preflight warm-up by simultaneously contracting their major upstroke and downstroke muscles. Yet something about the cuculiinid behavior was striking: some of them began to shiver at much lower temperatures than other moths of the same size.

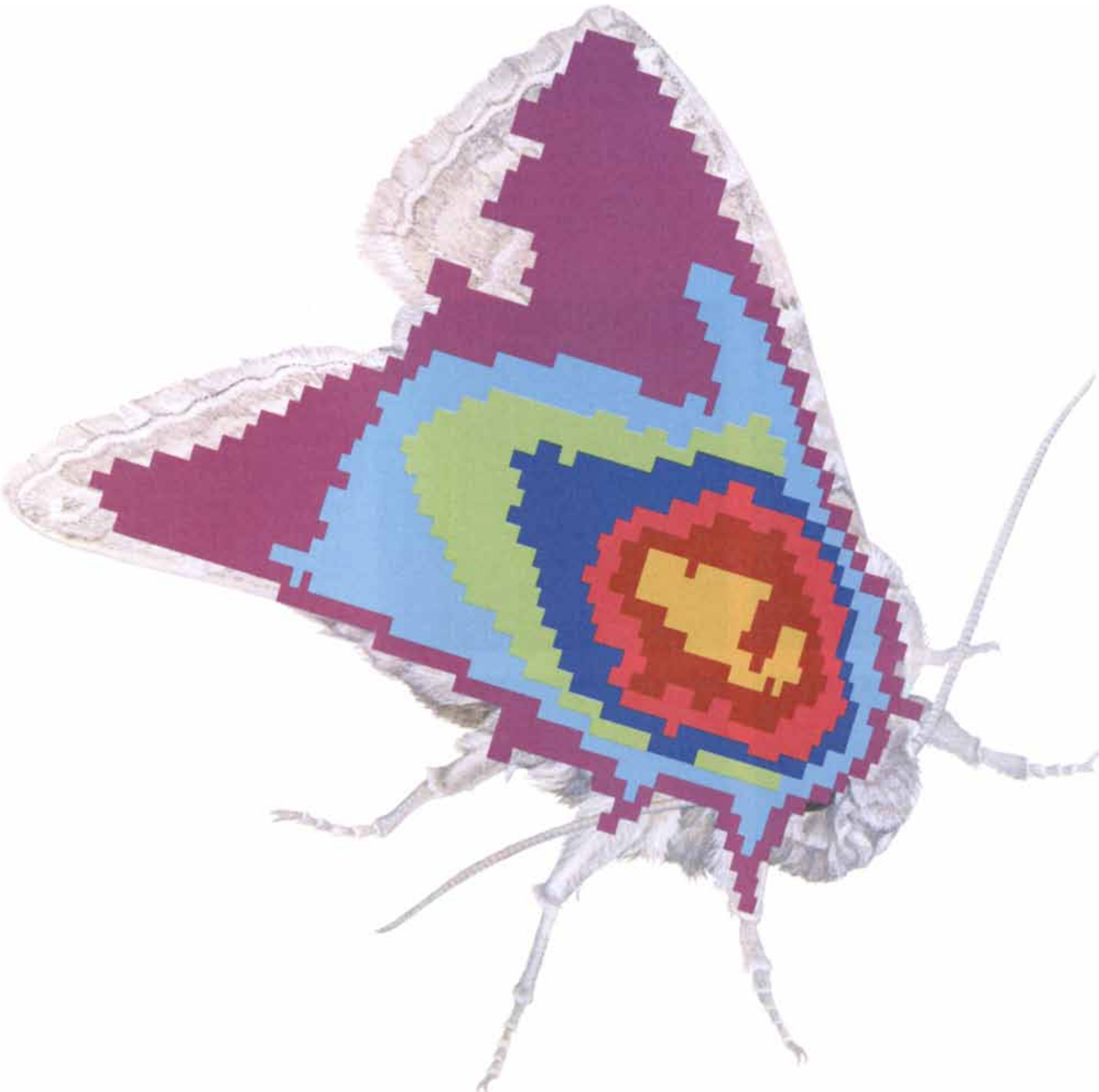
Winter cuculiinids generally become active only when the air temperature exceeds zero degrees C., but sometimes they initiate shivering at temperatures as low as −2 degrees. No other moths are known to shiver until the air reaches at least 10 degrees. According to Harald E. Esch of the University of Notre Dame, the winter cuculiinids can activate their central nervous system, and hence their shivering response, at unusually low temperatures. How they do so is still something of a mystery.

Heating up to 30 degrees C. from a resting temperature of zero degrees or below can drain a lot of precious energy. I therefore wondered whether the winter moths had a high metabolic rate, enabling them to generate heat more quickly and efficiently than other moths. They do not. As measured by oxygen consumption, their metabolic rates at rest and during shivering and flight were roughly the same as those found by other investigators for many moth species with a similar body mass.

The winter moths, in fact, pay a price in time and energy for heating themselves at low temperatures. When they warm up in an air temperature of zero degrees C., much of the heat they generate diffuses into the air, and in many instances they have to shiver for more than half an hour to reach a thoracic temperature of 30 degrees. During flight the large differential between the air temperature and the insects' body temperature results in accelerated heat loss, forcing the moths to repeatedly stop, shiver and warm up again. In contrast, if the insects wait to shiver until the air temperature is nearly 10 degrees, they can warm up faster. Moreover, the heat they generate is sufficient to maintain nonstop flight at a thoracic temperature of from 30 to 35 degrees.

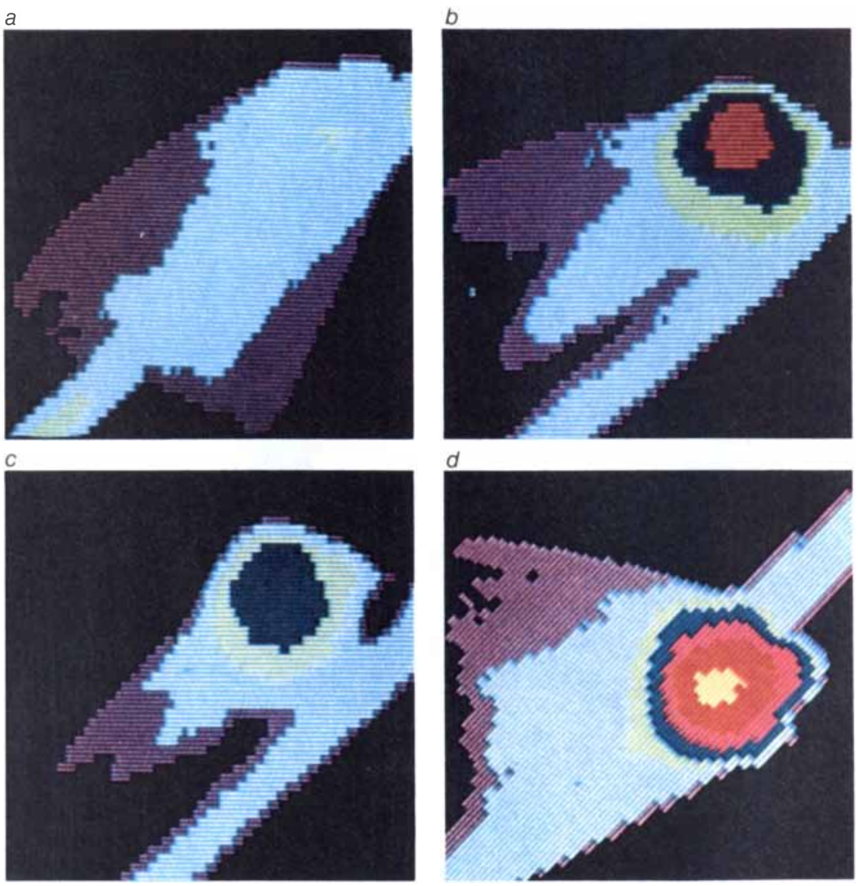
Lacking "low cost" ways of producing heat, the moths seem to be selective about the conditions under which they shiver. Intuition suggests that the animals should shiver when they feed, in order to be ready for a fast getaway from predators. Yet they sometimes do not do so. (Bats and birds may no longer patrol the skies in winter, but shrews, squirrels and perhaps deer mice probably pose some danger.) Moreover, the insects do not heat themselves merely to keep warm; if they do not need to fly, they neither warm up nor resist cooling after a flight. In fact, the lower the air temperature is, the less likely the moths are to shiver. At from five to eight degrees C. only half (49 percent) of the moths in one study shivered as they lapped up diluted honey smeared on trees. In contrast, most of the moths (90 percent) shivered at approximately 17 degrees, the highest temperature at which I observed them in the field.

This apparent emphasis on energy conservation at the expense of predator avoidance makes sense if one examines the costs of resisting passive cooling. At an air temperature of zero degrees C. a moth weighing .1 gram (having a thorax weighing .04 gram)



WINTER MOTH of the Noctuidae family is portrayed with a superposed infrared map to show the distribution of heat immediately after flight. Yellow indicates the highest temperature, followed in sequence by red, pink, dark blue, green, light blue and purple. In order to fly, the winter noctuids must keep their wing muscles (*yellow, red and pink regions*) heated to about 30 degrees Celsius.

TWO WINTER NOCTUIDS, *Eupsilia* (*left*) and *Lithophane* (*right*), eat honey spread on a tree trunk as bait. Some 50 noctuid species, all from the Cuculiinae subfamily, are active in northern winters. The winter cuculiinids resemble many summer noctuids anatomically but can warm their flight muscles at lower air temperatures. The cuculiinids are also better able to retain heat in their thorax, the region that houses the wing muscles.



INFRARED PHOTOGRAPHS of *Eupsilia* capture the moth in successive stages of warming up for flight. Like many moths, the cuculiinids heat up by shivering: simultaneously contracting their upstroke and downstroke muscles. In *a* and *d* the moth, perched on a stick, is seen from above, in *b* and *c* from the side. The thorax is the round area that first appears in *b* (*dark blue*). As in the illustration on the preceding page, yellow represents from 26.6 to 30.9 degrees C.; red, 24.8 to 26.5; pink, 22.4 to 24.7; dark blue, 19.6 to 22.3; green, 17.2 to 19.5; light blue, 14.0 to 17.1, and purple, 11.0 to 13.9 degrees.

would have to counteract a postflight cooling rate of 13 degrees per minute to maintain a 30-degree difference between its body and the environment. In order to do so it would expend .42 calorie per minute.

If the moth in the example above filled itself to capacity on sugar-maple sap, obtaining four milligrams of sugar, its shivering would exhaust the contents of its stomach in just 35.2 minutes (each milligram of sugar provides 3.7 calories). At an air temperature of 15 degrees, on the other hand, the moth would burn calories much more slowly. In fact, it could maintain a 30-degree thoracic temperature for twice as long.

Because the moths do not have highly specialized mechanisms for producing extra warmth, I suspected that they must have an effective way of retaining heat. Actually they have several ways. Insulation can certainly retard heat loss, and the moths are well insulated by a coat of dense pile. (The pile is a derivative of the scales that give butterflies their beauty, and it explains why the moths are often called millers: the modified scales, which rub off easily, are whitish and fluffy, much like the flour that covers a miller.)

To determine exactly how well the scales facilitate heat retention, I measured the cooling rates of both pile-covered and depilated moths after they were heated and then exposed to varying air speeds in a wind tunnel. At air speeds of seven meters per second, which approximate those of flight, the unshorn moths cooled approximately twice as slowly as the naked ones.

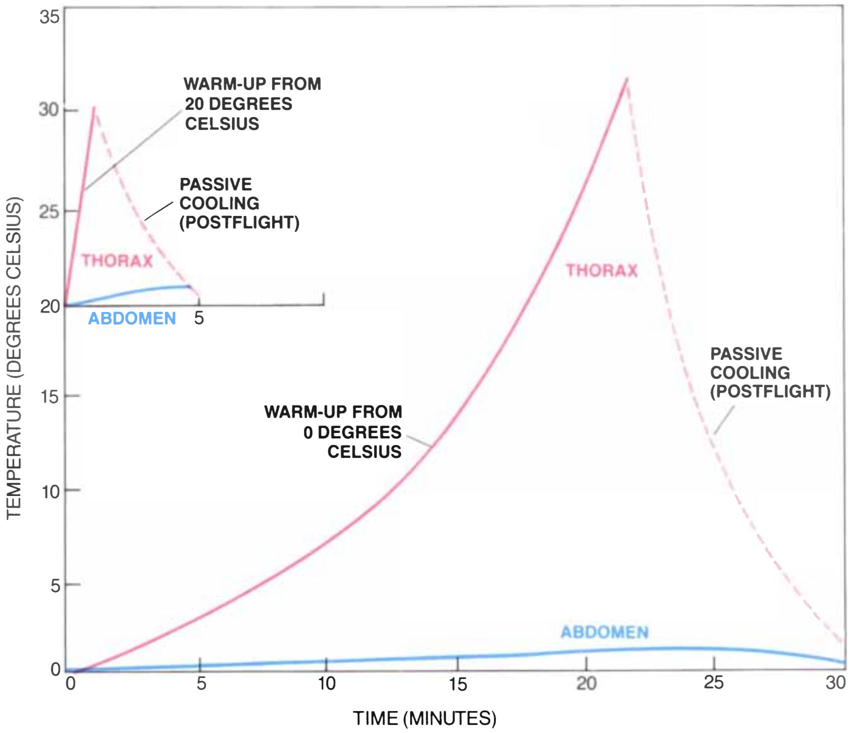
Pile obviously helps the animals to store heat; indeed, it is absolutely essential to flight in the winter. Nevertheless, some summer moths of similar size, such as the tent-caterpillar moth *Malacasoma americanum,* have comparable insulation, to aid them when summer nights turn chilly. The insulation alone, then, does not explain why the winter cuculiinids withstand the cold better than other moths.

Like the insulation, an ability to prevent heat from diffusing out of the thorax into colder areas of the body would also help the moth to retain crucial thoracic warmth. Indeed, to varying degrees all endothermic insects examined to date have such an ability, including dragonflies, bumblebees, honey bees and many large moths. At low air temperatures, when they need to conserve thoracic heat, the animals retard heat flow to the head and the abdomen and virtually eliminate leakage to other extremities, such as the legs and wings. The winter cuculiinids do the same, but compared with other moths they lose even less heat to the abdomen.

Determining the thoracic and abdominal temperatures of cuculiinids required inserting fine thermocouple probes into them before they warmed up. During the preflight warm-up the abdomen remained within .4 degree C. of the air temperature. Indeed, the abdominal temperature increased an average of only two degrees even during flight, whereas the thoracic temperature increased as much as 35 degrees.

In collaboration with George R. Silver, formerly of the Cold Research Division of the U.S. Army Research Institute of Environmental Medicine in Natick, Mass., I also photographed the insects with an infrared camera. The camera records heat emission rather than external details. Our pictures confirmed that the legs, wings and abdomen of the winter moths gain little or no heat during warm-up, flight and postflight cooling.

How can a moth maintain a difference of more than 30 degrees C. between the thorax and the abdomen, trunk divisions separated by only one or two millimeters? Oddly, part of the answer lies in the anatomy of the moths' ears. The eardrums of noctuids lie behind the thorax, where they are enclosed by air chambers that happen to be almost perfect heat insulators. (No one knows whether the ears of winter moths still function as they once did to detect the ultrasounds emitted by bats.) Moreover, these chambers sit next to abdominal air sacs, which provide added insulation.



THORACIC AND ABDOMINAL TEMPERATURES of *Eupsilia* are plotted over time as the winter cuculiinid prepares itself for flight. When the air temperature is 20 degrees C. (*inset at left*), the moth can warm its thorax to 30 degrees within 1.5 minutes, whereas it must shiver for 22 minutes when the air is at zero degrees (*right*). In order to conserve energy, winter cuculiinids usually avoid heating up for flight when the air temperature is near zero. They also save energy by preventing heat from leaking into the abdomen (*blue*), which always remains within a few degrees of the air temperature. (The moths were held in place by a tether; they cooled rapidly when they were unable to take off.)

The vascular system also helps to prevent heat loss from the thorax. The blood, which transports nourishment stored in the abdomen, could potentially nullify the heat-sequestering ability of the air sacs. It flows in a single vessel from the abdomen (where the vessel is called the heart) through the thorax to the head [*see illustration on page 109*]. From the head it empties into the surrounding tissue, eventually percolating back to the abdomen.

In theory, blood returning to the abdomen could carry heat away from the thorax. In practice, one section of the circulatory system in the abdomen and one section in the thorax act as countercurrent heat exchangers: biological structures that in this instance recapture heat before it fully escapes from the thorax. In a countercurrent heat exchanger two fluids (or gases) contained within separate but adjacent channels flow in opposite directions. If the fluid in one channel is warmer than the fluid in the other channel, heat moves from the warmer substance into the cooler one.

The cuculiinid's abdominal heat exchanger lies under the moth's air sacs. It consists of both the blood vessel transporting cool blood from the abdomen to the thorax and a narrow region of the tissue surrounding the vessel. Blood that has been heated in the thorax flows through this tissue to the abdomen, which means that it moves in the opposite (or counter) direction from the vascular blood. Heat from the tissue therefore diffuses into the cooler blood that is flowing into the thorax.

As the blood vessel leaving the abdomen enters the thorax, the tube becomes the aorta—and the cuculiinids' second heat exchanger. Once inside the thorax, the vessel forms an inverted U, in which the two arms are pressed close together. First the vessel travels upward to the top of the thorax; then it curves sharply downward (before eventually turning away to the head). The blood flowing into the thorax from the abdomen is initially cooler than the thorax, but it becomes heated as it travels. The blood in the descending loop is therefore warmer than the blood in the ascending part. As a result heat returns to the ascending loop instead of traveling to the head with the flowing blood.

An ideal way to evaluate the effectiveness of a heat exchanger is to alter the configuration of the blood vessel. For instance, one might separate the ascending and descending parts of the thoracic heat exchanger and predict that heat from the descending part would then be lost to the head. Unfortunately it is next to impossible to do such surgery in a tiny moth without interfering with many things, including the rate of blood flow, which can in turn affect the movement of heat throughout the body.

It is possible, however, to judge the value of the winter cuculiinid's heat exchangers by comparing its vascular system with that of other moths, such as sphinx moths (Sphingidae) and giant silk moths (Saturniidae), both of which are large-bodied animals found particularly in the Tropics. In these insects the aorta is arranged to form a "cooling coil" instead of a heat exchanger. The descending part is greatly elongated and loops away from the ascending part. Rather than returning heat to the other part of the aorta, the looping vessel retains the heat and transports it away from the thorax.

The differing physiologies appear to produce markedly different effects. Sphinx moths and giant silk moths are up to 60 times as large (in mass) as the cuculiinids and therefore might be expected to overheat much more readily. On the contrary, they can dissipate excess heat to the head and the abdomen and from there to the air. In fact, they often fly at air temperatures higher than 30 degrees C. The cuculiinids, in contrast, never "dump" their excess warmth. Regardless of their very small size, they stop flying because of heat prostration when the air temperature approaches 20 degrees. Their extremely efficient heat-retention mechanism is apparently bought at a price, but one they seldom if ever have to pay.

Although the winter moths' heat exchangers effectively retain thoracic heat, the insects' circulatory system, like their pile covering, is not totally unique. Indeed, it is similar to that of many small-bodied summer moths. For instance, in the aorta of the tent-caterpillar moth, the descending loop is close to the ascending one, but the two are not pressed together. This small difference nonetheless appears to influence heat retention. The tent-caterpillar moth, which in flight maintains the same thoracic temperature as the winter cuculiinids, has a modest ability to dump heat. It can therefore fly in slightly warmer weather than the winter moths can, but it is not able to fly at low air temperatures.

Producing and retaining the heat needed for flight is just one part of the solution to winter survival. Because winter moths actually spend at least 99 percent of the time cooled down and in torpor, they also need some way to avoid freezing, or solidifying, when they rest in wait for a suitably "warm" night.

The immature stages of many summer insects survive the cold by producing biological antifreezes. I wondered if the winter moths did so too. By storing moths in a refrigerator for an average of three weeks, John G. Duman of the University of Notre Dame and I determined the standard freezing point of the insects' blood; that is, the freezing point when internal ice crystals are present. Even tiny crystals promote freezing because they provide a surface for the attachment of nearby water molecules.

The moths froze at from −1 to −2 degrees C., close to the freezing point of summer-adapted insects. The blood of moths that were freshly caught in the field had almost the same freezing point. These observations suggested that the cuculiinids produce little if any antifreeze; if they had such a substance in the blood, they would not freeze until the blood reached a much lower temperature.

It was still possible that the moths had a special ability to be supercooled (to remain unfrozen at temperatures below the standard freezing point) if they prevented an initial "seed" ice crystal from forming in or entering the body. The limit of supercooling is defined as the moment at which "flashing," or rapid ice-crystal growth, suddenly freezes the supercooled animal. This moment was readily identified by a quick, transient temperature rise in the moths; as water molecules join a spontaneously propagating ice crystal during flashing, they emit heat, which can cause a temperature rise of a few degrees. The flashing point is also the moment of death for the winter cuculiinids, because none survive freezing.

In moths that were chilled very slowly in an ice-free environment the supercooling limit varied markedly, ranging from −4 to −22 degrees C. Moreover, the moment of freezing did not cluster around any one temperature. Such variation suggests that supercooling is a random phenomenon, not an adaptive trait, in the winter moths. Indeed, the supercooling ability of the winter moths is similar to that of many summer moths that never encounter freezing temperatures.

PILE COAT on a winter cuculiinid facilitates heat retention. Insulation helps to ensure the survival of winter moths but is not unique to them; many summer moths have such fur.

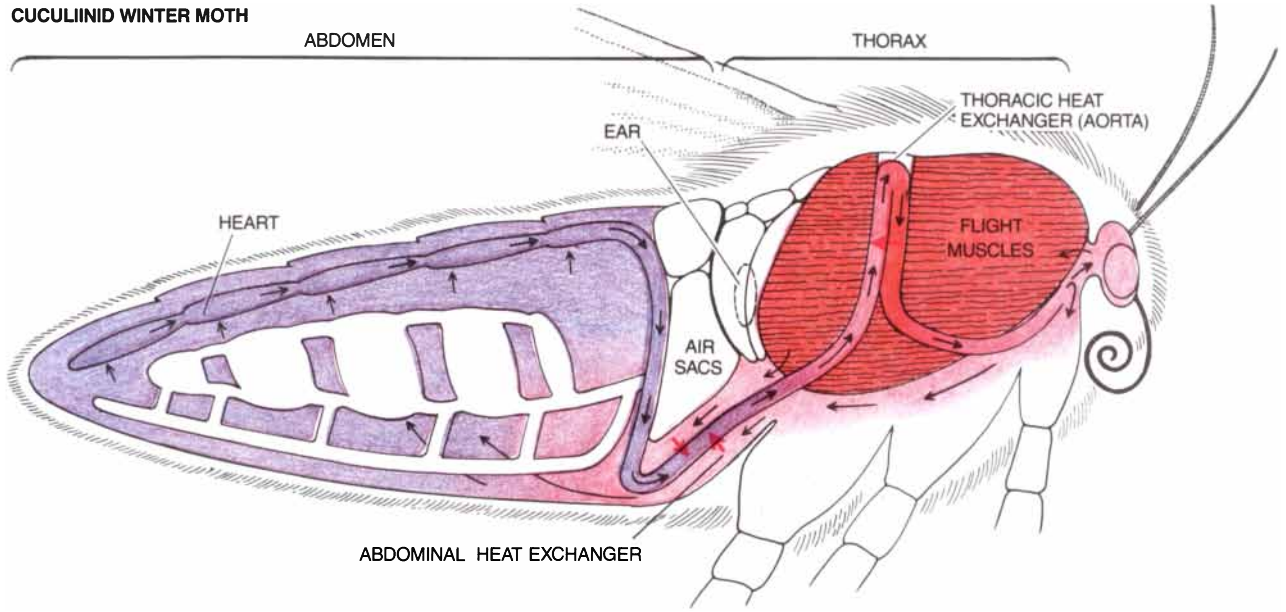
Although supercooling is probably not a specific adaptation for winter survival, any ability to cool below the standard freezing point may enable insects to survive low air temperatures, provided they stay in a dry place away from ice that might enter the body and serve as a seed crystal. A small, dry cavity in the ground, in a rotting log or under bark often forms an ideal habitat for hibernating insects. Where do the winter moths go to escape ice and sharp declines in their temperature?

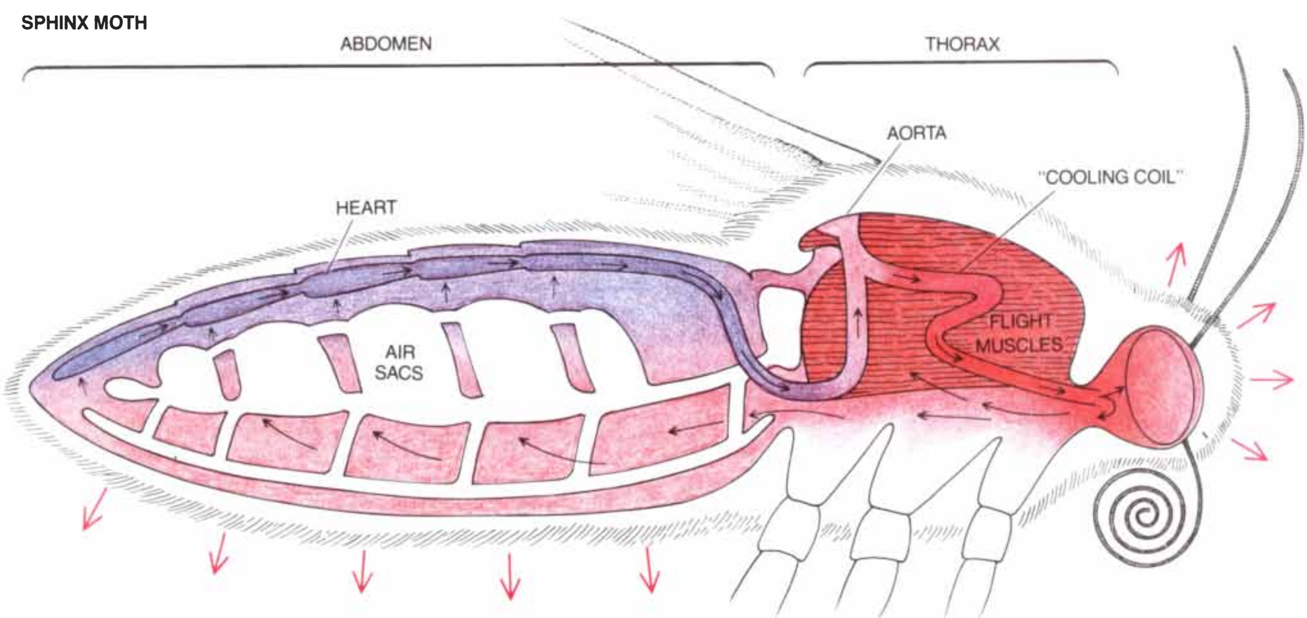
An observer who knew that coloration often provides protection might well suppose the moths rest on trees. Why else would they, in common with their summer-adapted relatives, be colored in ways that provide camouflage against tree trunks? There are, for instance, white- and cream-colored moths that are nearly invisible on birch; brown moths that blend with fir or spruce; black varieties that almost disappear against ash bark; gray ones that match beech or elm, and even some moths that are peppered to resemble lichens.

I discovered the moths' resting place by first building a large outdoor enclosure and standing in it the trunks of pine, spruce, birch, beech, ash, maple, elm and cherry; a leafy layer covered the ground. I then released 173 moths of varying colors one evening and searched for them the next morning. Most of the recaptured moths had crawled under the leaf litter or into curled leaves lying on the ground.

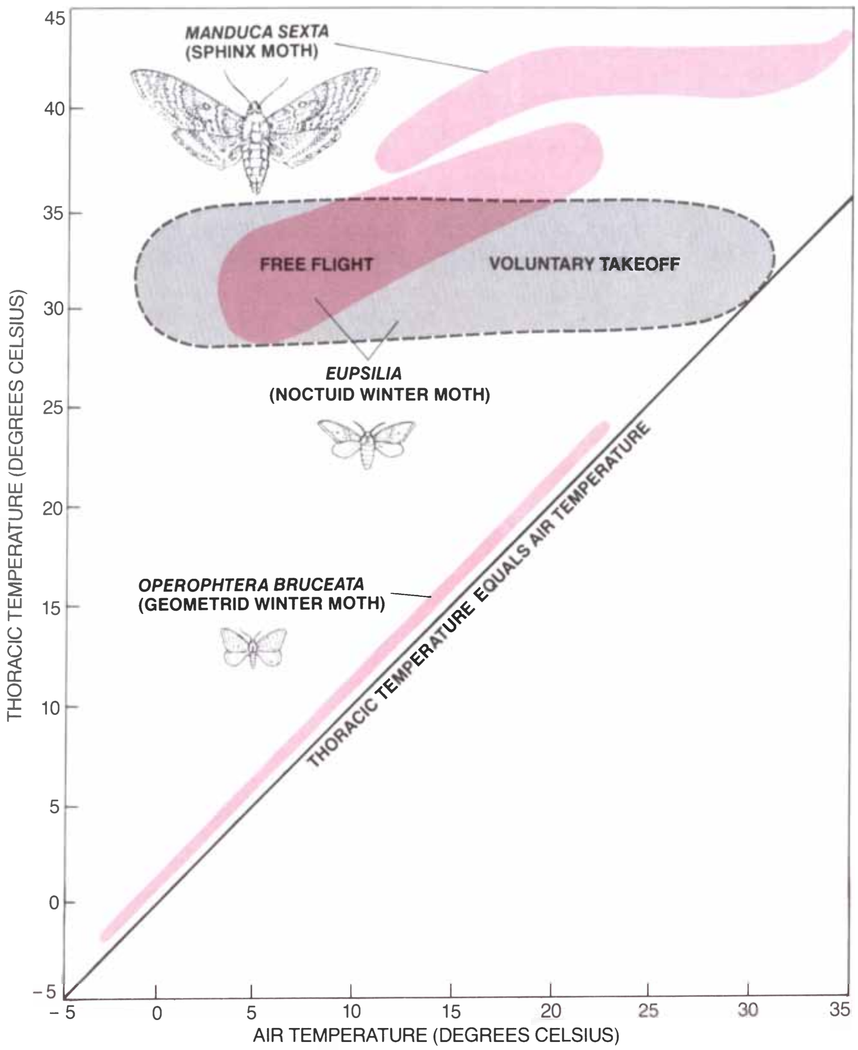
Leaves provide excellent insulation from the cold. In late winter in Vermont I measured temperatures of no less than −2 degrees C. under the ground cover even when the air temperatures dipped below −30 degrees. Such a makeshift shelter is often covered with snow, which provides an added but perhaps unnecessary barrier against the cold. Dale F. Schweitzer of Yale University has shown that the fallen leaves can be insulation enough, at least at temperatures as low as −23 degrees.

Moths that hide under leaves should not need camouflage. The explanation for their coloring may be found in their evolutionary history. The fact that the external and physiological

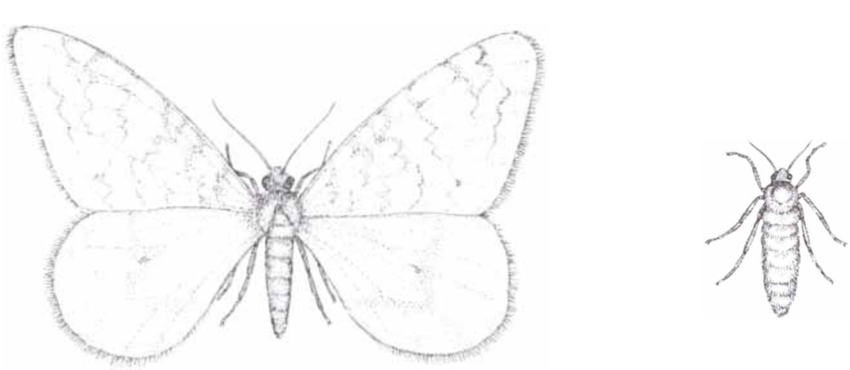




ANATOMY of a typical winter cuculiinid (*top*) differs from that of a summer moth (*bottom*) in ways that enable the cuculiinid to be active in the cold. The air sacs of the winter moth isolate the thorax, sequestering heat there. The circulatory system also preserves thoracic heat. In all moths the blood (*black arrows*) flows in a single vessel from the abdomen through the thorax to the head, warming along the way; on the return journey it percolates through tissue. The circulatory system of the winter moth includes an abdominal and a thoracic countercurrent heat exchanger. In the abdominal heat exchanger blood flowing between the heart and the aorta is cool (*blue*), whereas blood flowing in the opposite (counter) direction, through the adjacent tissue, is warmer (*red*); heat therefore passes (*red arrows*) from the tissue into the vessel, and then into the thorax. The thoracic heat exchanger is the aorta. Because the blood in the descending part of the loop is warmer than the blood in the ascending part, heat moves into the ascending segment. In the large summer moth the descending part of the aorta is separated from the rising segment, shunting heat away.



THORACIC AND AIR TEMPERATURES at which a summer moth engages in continuous flight are compared with those of representative winter moths. *Manduca sexta,* a summer sphinx moth, must heat its thorax to about 40 degrees C. before flight; it can remain aloft in air temperatures as warm as 35 degrees C. Like many other summer moths, it cannot initiate flight until the air reaches about 12 degrees. *Eupsilia* flies at a lower thoracic temperature. It can take off when the air temperature is in the range of −2 to about 30 degrees (*gray*), although it cannot sustain flight at the extremes of that range. When the air is cooler than five degrees, the moth stops flight repeatedly to shiver. When the air is warmer than about 20 degrees, the animal overheats. *Operophtera bruceata,* one of the few winter-adapted members of the Geomitridae family, can fly in air temperatures ranging from −3 to 25 degrees and does not have to heat its thorax much in order to do so. It overheats at air (and also thoracic) temperatures close to 25 degrees.



MALE *OPEROPHTERA* (*left*) has large wings, a feature that helps to explain why it can fly without significantly warming its wing muscles. In contrast to the winter cuculiinids, which beat their wings approximately 60 times per second, male *Operophtera* moths beat their wings as few as two times per second. The female of the species (*right*) does not fly.

features of the winter moths are quite similar to those of summer moths suggests that the winter-adapted variants evolved from ancestors that were active in the summer. If that is the case, the winter moths probably did perch on trees at one time. After they adopted new habits their color became a neutral trait and was not altered. Assuming that this hypothesis is correct, it suggests that a "reversed" life cycle evolved independently many times. Indeed, according to John G. Franclemont of Cornell University, taxonomic findings also suggest that the winter cuculiinids are polyphyletic (descendants of more than one ancestral line) and have all developed the same strategy for winter survival.

The moths frequently seek shelter to avoid freezing when they rest, but at other times they may optimize their energy balance by avoiding too cozy a hiding place. Indeed, the lower they keep their body temperature while resting (short of freezing), the longer they can make their energy reserves last. The reason is that the metabolism slows in cold weather. For example, on the basis of measurements of energy metabolism in resting moths I calculate that a moth weighing .1 gram and filled with six grams of sugar from sap can rest for 193 days at an air (and body) temperature of −3 degrees C. At three degrees higher, or zero degrees, the fuel would last for only 24 days, and at 10 degrees the reserve would be exhausted in just 11 days. Whether or not the moths actually attempt to remain at the lowest temperature possible is not yet known. To do so would entail an important risk: if the insects tried to sit still in an exposed place, they might become too cold and freeze.

The accumulated evidence suggests that the moths do not have any highly specialized adaptations for resisting the cold, just as they have no unique adaptations for generating heat. On the other hand, their behavioral adaptation of seeking shelter in the leaves serves them nicely. It allows them to be flexible. They can become active the moment the air is warm enough for efficient flight, and yet they can also find shelter within seconds if temperatures plummet dangerously during the same evening. Insects with adaptations enabling them to survive freezing by perfusing themselves with antifreezes would require a significantly longer lead time to become fully active—time that winter evenings may not provide. In high concentrations antifreezes (primarily alcohols) are toxic and temporarily leave animals comatose; the substances are eventually converted into less toxic chemicals, but that happens slowly, particularly when the temperature of an animal is very low.

Winter cuculiinids seem to be well adapted to cold weather, but it would be a mistake to conclude that their characteristics are the only ones possible for winter activity. On frosty nights in New England one can see the males of the species *Operophtera bruceata* sailing through the forest in search of the flightless, sluglike females. The moths, which also fly on sunny days, are even active at temperatures as low as −3 degrees C. and during mild snowstorms in November. (In late November, before they disappear entirely, the males fly only at about noon on sunny days.)

Operophtera, which is one of a few species in the Geometridae family that have adapted to winter, neither basks (a typical warming behavior of day-flying insects) nor shivers, nor does it have the insulation found in cuculiinids. Instead the males are able to function at an extremely low body temperature; they are the only moths that routinely fly with a muscle temperature close to zero. Spared from the need to warm up before takeoff, *Operophtera* saves the considerable energy that might otherwise go toward shivering.

Oversized wings and a low body weight contribute to such energy conservation by enabling the insects to remain airborne with a wingbeat frequency as low as from two to four beats per second—much lower than the more than 60 beats per second required by the cuculiinids. In common with many winter-adapted insects, the adults do not eat; in fact, they no longer have a digestive tract. (All the energy they use is accumulated and stored at the larval stage.) It is impossible to ascertain cause and effect, but the loss of the need to carry a digestive tract probably reduced the energy needed for flight.

Geometrids that live near equatorial lowlands have physical characteristics similar to those of *Operophtera,* but they are less exaggerated. As is true of the cuculiinid winter moths, *Operophtera's* design for cold-weather activity appears to be modeled on a preexisting form, albeit one quite different from that of the noctuids.

The remarkable ability of certain noctuids and geometrids to be active in the winter highlights the way slight evolutionary alterations in anatomy, physiology and behavior can add up to success in a new environment. The winter moths are still quite similar to their close relatives but, taken together, their small differences adapt the insects to winter living.

MESOLITHIC MOTHER AND CHILD were buried together in a grave found with other graves at a site near the modern Danish town of Vedbaek. The woman was a young adult, her child newborn. The object among the child's bones is a flint knife. The teeth around the mother's skull are from red deer; they were attached to a skin garment. (The round stone on the other side of the skull is the result of later burrowing by animals.) The cemetery, called Henriksholm-Bøgebakken, dates from about 4000 B.C. Its excavation, begun in 1975, focused renewed interest on the Vedbaek area, which in Mesolithic times was an inlet of the sea.

A Mesolithic Camp in Denmark

On a small rise that was once an island an unusual excavation has yielded clues to the rich foraging culture that flourished on the coasts of northern Europe during the "Middle Stone Age"

by T. Douglas Price and Erik Brinch Petersen

What can be said about a group that lives by foraging? Until recently the standard answer was that such a group is small, poor and moves more or less continuously in search of food. From this picture it follows that the group's social organization is rudimentary. That answer, which fits many of the foraging groups existing today, is being changed by new findings from a period that ended some 5,000 years ago: the Mesolithic. It has become increasingly clear that during the Mesolithic period (the "Middle Stone Age") there existed foraging groups that were relatively large, affluent and often sedentary. Furthermore, although their social organization is difficult to reconstruct, it seems likely that some of these groups were characterized by considerable social complexity.

Nowhere are these developments clearer than in southern Scandinavia, where a combination of favorable circumstances has made the archaeological record of Mesolithic times particularly rich and intelligible. Southern Scandinavia (including northern Germany, Denmark and southern Sweden) remained largely uninhabited until about 12,000 years ago, when, as the Pleistocene ice sheets retreated, the area was penetrated by bands of reindeer hunters. Later the reindeer hunters were replaced by a series of foraging cultures typified by increasing complexity and increasing dependence on marine resources. Fortunately evidence of this rich cultural sequence has been well preserved in the many bogs that formed in the region during and after the Mesolithic period.

Partly as a result of the excellent preservation of a wealth of cultural artifacts, archaeology constitutes a significant aspect of the national tradition of Denmark and Sweden. One area of Denmark where excavations have served to unearth a microcosm of Mesolithic life is around Vedbaek, a small town just north of Copenhagen on the island of Zealand. During Mesolithic times there was an inlet near the site of the modern town, and a series of excavations beginning 60 years ago have shown that the inlet was ringed by Mesolithic settlements. In 1975 the discovery of a significant Mesolithic cemetery spurred renewed interest in the area, and several major excavations have been carried out there in the past decade.

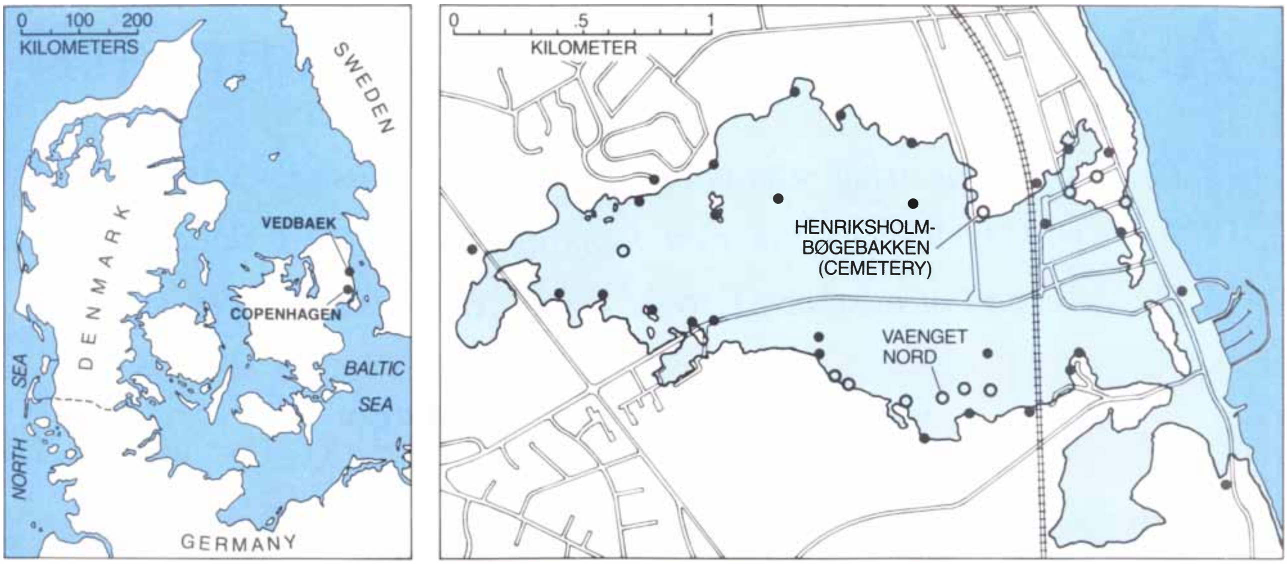
One of those excavations was our own work at Vaenget Nord. Now only a slight rise among birch trees in a meadow, Vaenget Nord was once a small island in the inlet. About 7,000 years ago the island was the site of a small camp that was probably seasonally occupied several times by a group of people whose descendants were to live in the area for the next 2,000 years. Beginning in 1980 and continuing until 1983 we undertook intensive excavations at Vaenget Nord. Making use of the technique called *décapage,* we peeled back the soil layer by layer, leaving the artifacts in place as we revealed the Mesolithic surface. That method is particularly effective for reconstructing the spatial organization of a prehistoric site. What we have learned about spatial organization has in turn contributed to our growing understanding of Mesolithic complexity.

During much of its prehistory northern Europe was covered by ice. At the end of the Pleistocene epoch, some 10,000 years ago, dramatic changes in climate and environment took place as a result of a global warming. As the Pleistocene ice retreated it left new sediments along with many shallow depressions that became lakes and streams. The new landscape was initially occupied by tundra species such as dwarf birches, lichens, reindeer and horses. These initial colonizers were quickly followed by human hunters belonging to various cultures of the late Paleolithic period (the "Old Stone Age"). As temperatures rose, tundra gave way to open parklands of birch and eventually pine. The light forests were filled with new species such as aurochs (wild cattle) and moose, soon supplemented by wild pigs, red deer and roe deer.

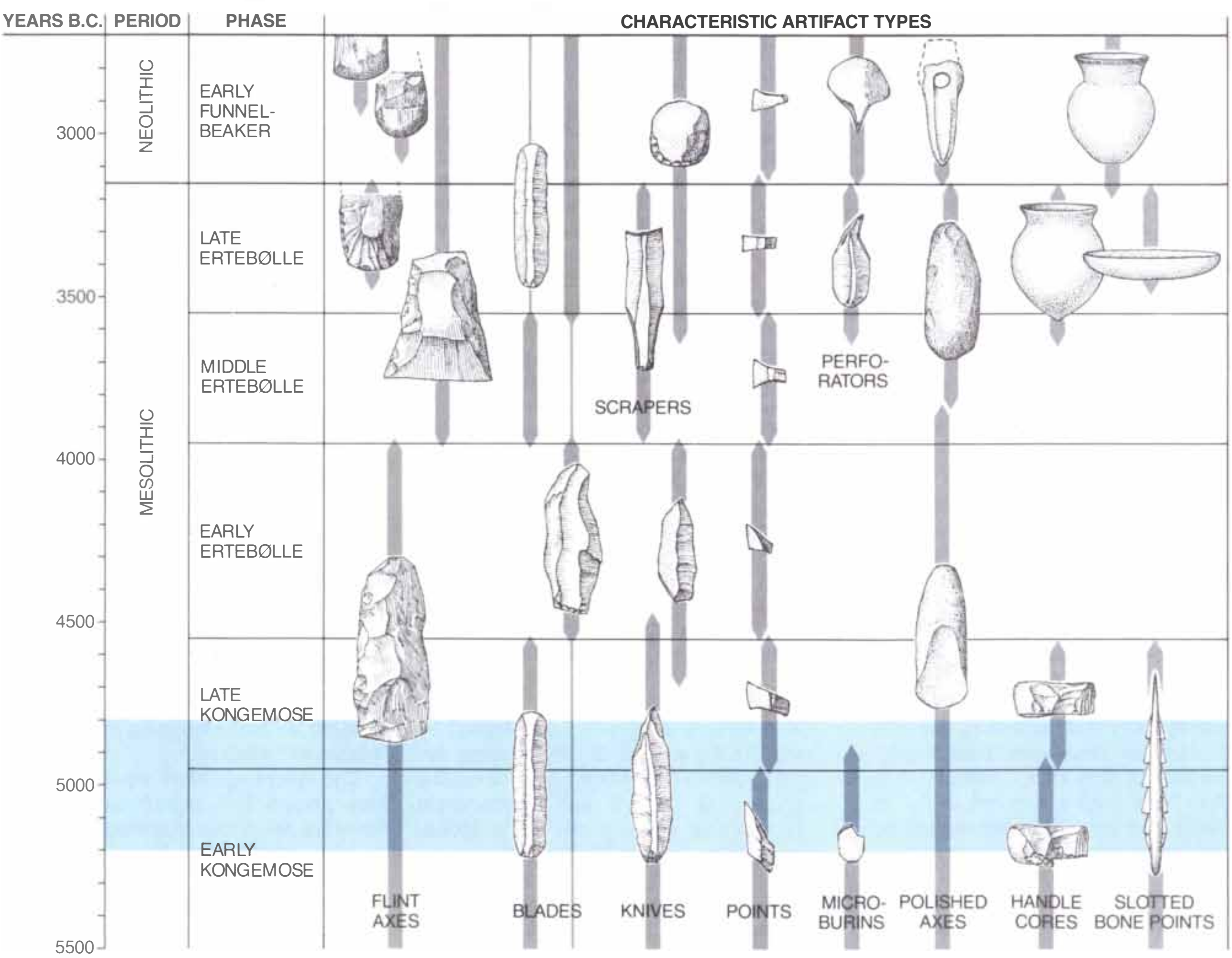
The first traces of Mesolithic culture appeared in southern Scandinavia at the beginning of the postglacial epoch. Archaeological remains dating to before 7000 B.C. are rare, but after that date foragers began to leave abundant traces of their presence. (The dates employed in this article are uncalibrated radiocarbon dates, which differ from calendrical dates by an amount that increases with the interval before 400 B.C. The uncalibrated dates can be adjusted to the calender by means of bristlecone pine trees of known age.)

By analyzing the various traces the foragers left, archaeologists have been able to divide the Mesolithic in southern Scandinavia into three main periods. These three chronological divisions are known as the Maglemose, the Kongemose and the Ertebølle, each named for the most notable of its characteristic sites. The first period, the Maglemose, lasted until about 5500 B.C. Maglemosian hunters are known mainly from small, short-term summer and fall camps on inland lakes. The focus of these encampments seems to have been fishing, complemented by some hunting and the collection of hazelnuts.

The warming trend of the postglacial period had significant implications for the physical setting of human culture in southern Scandinavia. As the continental ice sheets melted, the seas rose. At the same time the landmasses, released from the weight of an ice blanket hundreds of meters thick, also began to rise. A kind of geologic race ensued between land and sea. Gradually the water outdistanced the land, and the sea rose until it had



VEDBAEK is on the Danish island of Zealand just north of Copenhagen (*left*). Today the area is a low-lying coastal valley; the current coastline and other features are shown by the light lines on the map at the right. In 4000 B.C. sea level was five meters higher than it is today (*dark line*). The resulting inlet was densely populated during Mesolithic and early Neolithic times. Of the many known Stone Age settlements around the inlet (*circles*), about a dozen have been excavated (*open circles*). From 1980 to 1983 the authors directed an excavation at Vaenget Nord, which in Mesolithic times was a small island 40 to 50 meters from shore.



MESOLITHIC CHRONOLOGY for Denmark can be constructed by examining changes in artifacts. Vaenget Nord was occupied from about 5200 to 4800 B.C. At the end of the Mesolithic ceramics appeared, probably adopted from farming groups to the south. The arrival of the type of pottery called funnel-necked beakers in about 3200 B.C. (along with domesticated plants and animals) marks the beginning of the Neolithic period. Peter Vang Petersen of the National Museum of Denmark conceived the diagram.

exceeded its current levels. By about 5000 B.C. southern Scandinavia was smaller than it is today, and many low-lying areas near the modern coast had been flooded.

The rich estuaries, inlets and islands formed by the rising sea were the focus of human settlement during the Kongemose and Ertebølle periods. Not only was the coast the site of settlement but also the sea became the main source of subsistence. Successful hunting on the many islands helped to gradually extirpate the bears, aurochs and moose. Subsistence activities turned to the wild pig, the red deer and most emphatically to the sea. Throughout the later Mesolithic there is a tendency for more marine species to appear in the diet and for those species to constitute a greater proportion of the diet.

By the late Mesolithic the coastal cultures included groups of sophisticated, marine-adapted hunters and fishermen. Along with large and exotic prey such as porpoises and whales, these marine-adapted groups exploited more prosaic resources of the sea such as fish and shellfish. In western Denmark remains of settlements are often found in association with huge mounds of shells from oysters, mussels, periwinkles and cardium (a type of scallop). Such mounds, known as kitchen middens (or *køkkenmødding* in Danish), were not the only signs of change. Settlements became larger, more elaborate and more sedentary. The increased permanence of settlement is suggested by the appearance of cemeteries.

EXCAVATION AT VAENGET NORD had a twofold plan: to expose broad areas on the former island and to cut deep trenches extending from its shore. One trench is shown in the foreground; the group in and near it is studying a section of the trench wall, aided by the matchsticks indicating the vertical arrangment of layers below the modern surface.

These developments can be seen quite clearly around the Vedbaek inlet, which is today on dry land a few hundred meters from the coast. The geologic history of the area has been carefully reconstructed by Charlie Christensen of the National Museum of Denmark. The Vedbaek area was given its present form by geologic deposits from the last Pleistocene ice sheets. The inlet itself was originally a valley formed by the runoff of glacial meltwater under the ice sheet. When the ice retreated, the valley held a freshwater lake-and-stream system. By about 5500 B.C. the rising postglacial seas had invaded the valley's seaward end, which soon became an inlet of the Øresund (the strait between the east coast of Zealand and the western shore of Sweden).

Seven thousand years ago the inlet offered a most propitious environment for human habitation. A diverse array of food species was available from sea and land. About 60 species of fish, reptiles, birds and mammals have been identified at the sites in the Vedbaek area by Kim Aaris-Sørensen of the Zoological Museum of the University of Copenhagen. The 60 species come from all segments of the local environment: the forest, streams, lakes, wetlands, the inlet, the sound and the sea. Forest animals are dominated by the red deer, the roe deer and the wild pig.

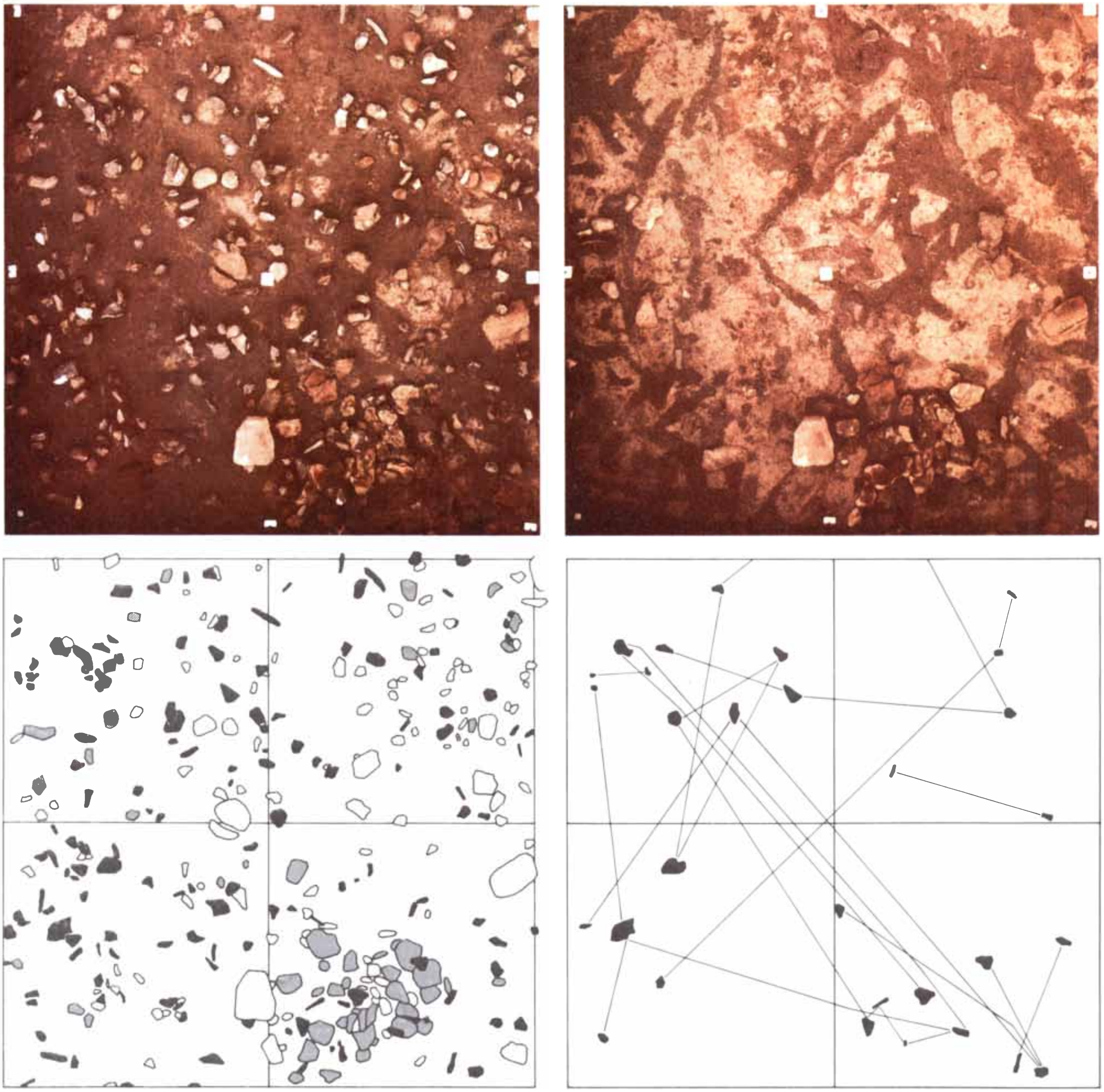
In spite of the diversity of the resource base, marine foods provided the major part of the diet. Measurement of two isotopes of carbon (whose proportions vary according to the source) in human bone suggests that the prehistoric inhabitants of the inlet depended as heavily on the sea as do the modern Eskimos of Greenland, about 75 percent of whose diet comes from marine sources. The significance of the marine environment is also attested to by the fact that all the Mesolithic settlements are adjacent to the prehistoric shoreline.

The richness of the environment around the inlet was reflected in a high density of human settlement. Forty or more Mesolithic sites have been identified at the edge of the inlet, many of them occupied repeatedly over a long period. Indeed, the work done so far indicates that the shores of the inlet were first occupied by Mesolithic settlers in about 5200 B.C. as the inlet was forming. Thereafter the inlet was occupied continuously by Mesolithic groups until the introduction of agriculture to the area in about 3200 B.C. marked the beginning of the Neolithic period (the "New Stone Age").

The repeated occupation of the sites around the inlet has provided a layered vertical arrangement of tools and debris corresponding to the chronological subdivisions of the Mesolithic. Such an arrangement, which archaeologists refer to as a vertical stratigraphy, can be employed for constructing a detailed chronology, and Peter Vang Petersen of the National Museum has done just that. Petersen's chronology divides the middle and later parts of the Mesolithic represented at Vedbaek into five phases, each of from 400 to 600 years' duration. Each phase is associated with distinctive types of artifacts, such as stone blades and points [*see bottom illustration on opposite page*].

Such a stratigraphy reveals some intriguing things about the relation between foraging and agriculture. Various artifacts found in the upper (later) Mesolithic layers show that the foragers had contact with farming groups over a period of perhaps 500 years before agriculture was adopted in Denmark. One sign of such contact is the appearance of pottery, generally considered part of Neolithic culture, in the latter part of the Mesolithic archaeological record. It would seem that the success of Mesolithic foraging cultures delayed the introduction of farming into southern Scandinavia for several centuries—during which farming was known but not adopted there as a way of life [see "Postglacial Foraging in the Forests of Europe," by Marek Zvelebil; SCIENTIFIC AMERICAN, May, 1986].

Although the Vedbaek area has been the subject of archaeological investigation for several decades, its significance was reemphasized in 1975 by the discovery of the Mesolithic cemetery, which has provided radiocarbon dates corresponding to about 4000 B.C. The cemetery contains the graves of at least 22 people, including eight adult males, an equal number of adult females and five infants. The burials were adorned in various ways. One infant was buried on the wing of a swan



DÉCAPAGE, the archaeological technique employed at Vaenget Nord, exposes broad areas of the former living surface; the four panels show some of the information that can be obtained from a square two meters on a side by means of this method. When the surface is first exposed, the artifacts are left in place to record their position (*upper left*). After most of the artifacts have been removed, the stains left by Mesolithic cooking and other activities (including the later action of moles) are revealed (*upper right*). Close analysis of the artifacts makes it possible to categorize them (*lower left*). Some of the stones were flint that had been worked (*black*); others were stones that had been cracked by fires (*gray*); still others were unmodified stone (*white*). Some bits of flint can be fitted together, which shows that they were originally part of a single piece of stone that was being worked (*lower right*).

next to his mother. Red-deer antler was placed with elderly people. Males were buried with flint knives, females often with jewelry made of shells and animal teeth. A few of these teeth—incisors of bears, aurochs and moose—come from animals that were extinct in the area at the time; they were probably obtained from central or northern Sweden or the continent of Europe by exchange.

The discovery of the cemetery stimulated renewed scientific interest in the Vedbaek area as the focus of a collaborative project aimed at describing and explaining changes in hunter-gatherer society in eastern Zealand from 5500 to 3000 B.C. The collaborating institutions included the Institute of Prehistoric Archaeology and the Zoological Museum of the University of Copenhagen, the Anthropological Laboratory of the University of Cambridge, the National Museum of Denmark and the departments of anthropology at the University of Wisconsin at Madison and the University of Winnipeg. The major excavations in the project have naturally been done at sites that were repeatedly occupied, because the cultural layer is thickest there. Repeated residence, however, tends to smear and obscure information on the horizontal arrangement of a settlement: the location of structures and the distribution of hearths, pits, artifacts and refuse.

Yet it seemed that the spatial arrangement of a settlement might offer substantial clues to the lives and social organization of its occupants. Accordingly, one aspect of the Vedbaek project was an effort to find sites with relatively unmixed remains and investigate their horizontal distribution. Several observations suggested Vaenget Nord might provide a good site for that kind of work. The slight rise that was the island of Vaenget Nord is today a minor undulation only 2.75 meters above sea level in a poorly drained meadow with birch woods. Preliminary work showed the site had been occupied during the early part of the Kongemose period. Because the island was flooded not long afterward, however, the period of potential occupation was brief. Indeed, test excavations showed that the number of artifacts per square meter was low compared with other sites in the area.

Major excavations at Vaenget Nord were begun in the summer of 1980 and continued until 1983. Teams of Danish and U.S. archaeologists collaborated in a twofold excavation plan based on what was known about the horizontal organization of many Mesolithic sites, where three separate zones can often be distinguished. The actual occupation zone is always adjacent to the shoreline. It contains hearths, pits and construction stone but little organic refuse (which is not well preserved in the open air). A refuse zone on the floor of the former inlet contains a variety of waste the inhabitants threw into the water, including bone, antler and flint. The refuse zone may also hold posts, fish traps and other wood objects. At some sites there is a strand zone between the occupation and refuse areas that has been cleared of artifacts by wave action.

The twofold excavation strategy at Vaenget Nord was intended to uncover both the occupation zone and the refuse zone in detail. Deep trenches were extended from the shore of the former island to expose the refuse zone and allow a careful study of the layers deposited in the former inlet. On the surface of the island large areas were exposed, the surface peeled back a layer at a time while the artifacts were left in place so that their position could be recorded. Because such *décapage* is time-consuming (and therefore expensive), it is rarely applied to large areas of a site. Yet it is one of the best ways to obtain detailed information about the horizontal arrangement of artifacts, and it is fortunate that the collaboration between our institutions made it possible to carry out the horizontal excavation on a fairly large scale.

Indeed, most of the surface of the former island was exposed. During its occupation the island was probably about 20 meters long and 10 meters wide; of the total of 506 square meters we investigated, 226 were exposed in the horizontal *décapage.* This peeling process exposed a cultural layer made up of ash, charcoal and other refuse that varied in thickness across the island. Under this layer the traces of human activities such as digging, fire building and setting stakes have been preserved as dark stains in the light sand of the island's surface. Our technique entailed the precise recording of the position of all these remains. In addition, all soil removed from the surface was sieved with water to obtain charcoal, bone, flint and plant remains that were too small to have been recorded in position.

The removal of the artifact-containing layer revealed many modifications of the surface, such as hearths and pits. These features were concentrated in the southwestern part of the island, the part nearest the mainland, which at the time of occupation was only 40 to 50 meters away. It seems clear that the landward side of the island constitutes the occupation zone. At the south and east margins of the occupation zone are areas where single large boulders have been fractured into fragments and laid down as a kind of small pavement. The bases of wood posts were also found scattered throughout the area of occupation. The posts, of birch or alder, were as much as 30 centimeters in diameter and sharpened to a fine point. Some of these, however, may be remnants of later structures associated with fishing.

The analysis of features and artifacts has enabled us to divide the occupation area into several subunits with different functions. In the western part of the island lies an area that appears to have been the habitation (or primary occupation) zone. Its most notable feature is a large, shallow depression that may have resulted from repeated trampling. A dense group of stake holes nearby may have continued around all or part of the depression. It seems likely that the depression formed the base of a dwelling of some kind. Little is known, however, about the details of Mesolithic dwellings, and so it is not possible to say with confidence whether it was a tent made of hides or a more substantial cabin of logs or peat.

In addition to the habitation area there were at least two other areas of specialized activity within the occupation zone, and the distribution of flint artifacts was of major significance in delineating them. The flint assemblage can be divided into the actual tools on the one hand and waste products, raw materials and other elements of production on the other. The waste, which archaeologists often refer to by the French word *débitage,* includes cores from which flakes and blades were made, core-shaping flakes and a great many small chips, or "shatter," resulting from the striking of cores. In some instances the pieces can be refitted, providing information about manufacture and about the discarding of the material.

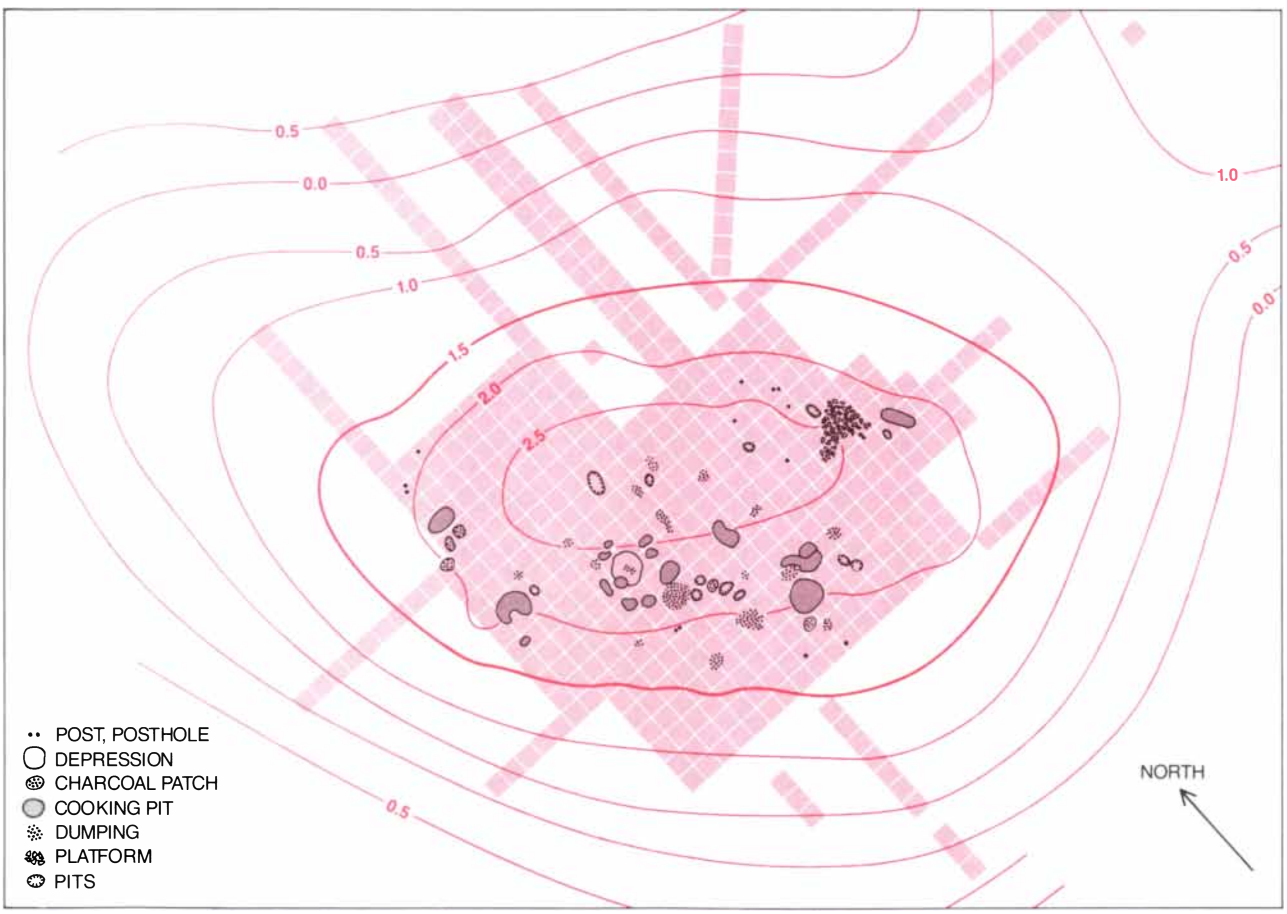
The tools themselves are of several types: axes, arrowheads, scrapers and burins (a tool with an edge resembling that of a chisel). The axes, weighing a kilogram or more, were made by flaking a heavy nodule of flint into an elongated implement with a broad, sharp edge. They were attached to long handles of elm or ash. The axes could be resharpened by removing a single large flake across the leading edge. Arrowheads generally have the shape of a rhombus (an equilateral parallelogram). Burins and scrapers were made from blades and flakes of flint by giving the stone the required edge. (Blades and flakes are distinguished by the fact that blades have a length of at least twice their width.)

The assembly of flint artifacts from Vaenget Nord is dominated by axes, arrowheads and burins; few scrapers are present. It has traditionally been assumed that burins were employed for engraving designs on such materials as bone, whereas scrapers were used on animal skins. To test these assumptions we submitted the flint artifacts from Vaenget Nord to Helle Juel Jensen of the Institute for Prehistoric Archaeology at the University of Århus. Jensen examined the edges under the microscope for the characteristic traces of wear that come from various materials [see "The Functions of Paleolithic Flint Tools," by Lawrence H. Keeley; SCIENTIFIC AMERICAN, November, 1977]. Jensen's investigation confirmed our assumptions about the uses to which the tools had been put: burins had been employed for shaping and shaving bone, the few scrapers present at the site had been used for working dry hides and the core axes exhibited heavy wood polish along the cutting edge.

With this information we were in a position to continue our mapping of the activity areas of the island. Most of the bone-working tools were found in the primary occupation area, and although projectile points were distributed throughout the site, their maximum density is also there, suggesting that the inhabitants made bone tools and hunting implements in or near their dwellings. Just to the northeast of the primary occupation zone, lies the area where most of the scrapers were found, along with truncated pieces of flint and unretouched blades. Few surface features are found there, and it seems probable that this was an area where hides were laid out on the ground and worked. Surrounding and overlapping the hide-working area is a zone that seems to have been the site of intensive flint manufacture; most of the refitted chips come from this area.

The overlapping hide- and flint-working areas (in addition to the primary occupation area) complete the list of the main components of the occupation zone. To the south of the occupation zone lies a zone of garbage dumping. The dumping area is characterized by a very thick, dark cultural layer containing an abundance of charcoal and fire-cracked stones but only a few dispersed flint flakes. This area is largely without surface features and the refuse layer is thickest next to the primary occupation zone.

To the north, on the opposite side of the island, is a second refuse zone that had two subsections, each with somewhat different contents. The presence of the first subsection was initially indicated by a high density of flint artifacts. Careful stratigraphic work showed that some of them had been dumped in the original refuse zone, whereas others had been redeposited later in an episode of erosion. Just to the north of that region we found a small area where bones and nutshells had survived both erosion and decomposition. Most of the shells are those of the hazelnut, which is known to have been a staple for some Mesolithic groups. Among the bones are those



SITE PLANS OF VAENGET NORD show artifacts and modifications of the living surface (*left*) along with specialized activity areas (*right*). In 5000 B.C. the outline of the island was defined by the 1.5-meter contour. A depression in the southwestern part of the island may have been the base of a dwelling. There are cooking pits and other pits as well as patches of charcoal, posts and

of several aquatic species, such as gar, mackerel, dogfish and even sting ray, in addition to the normal sylvan fauna.

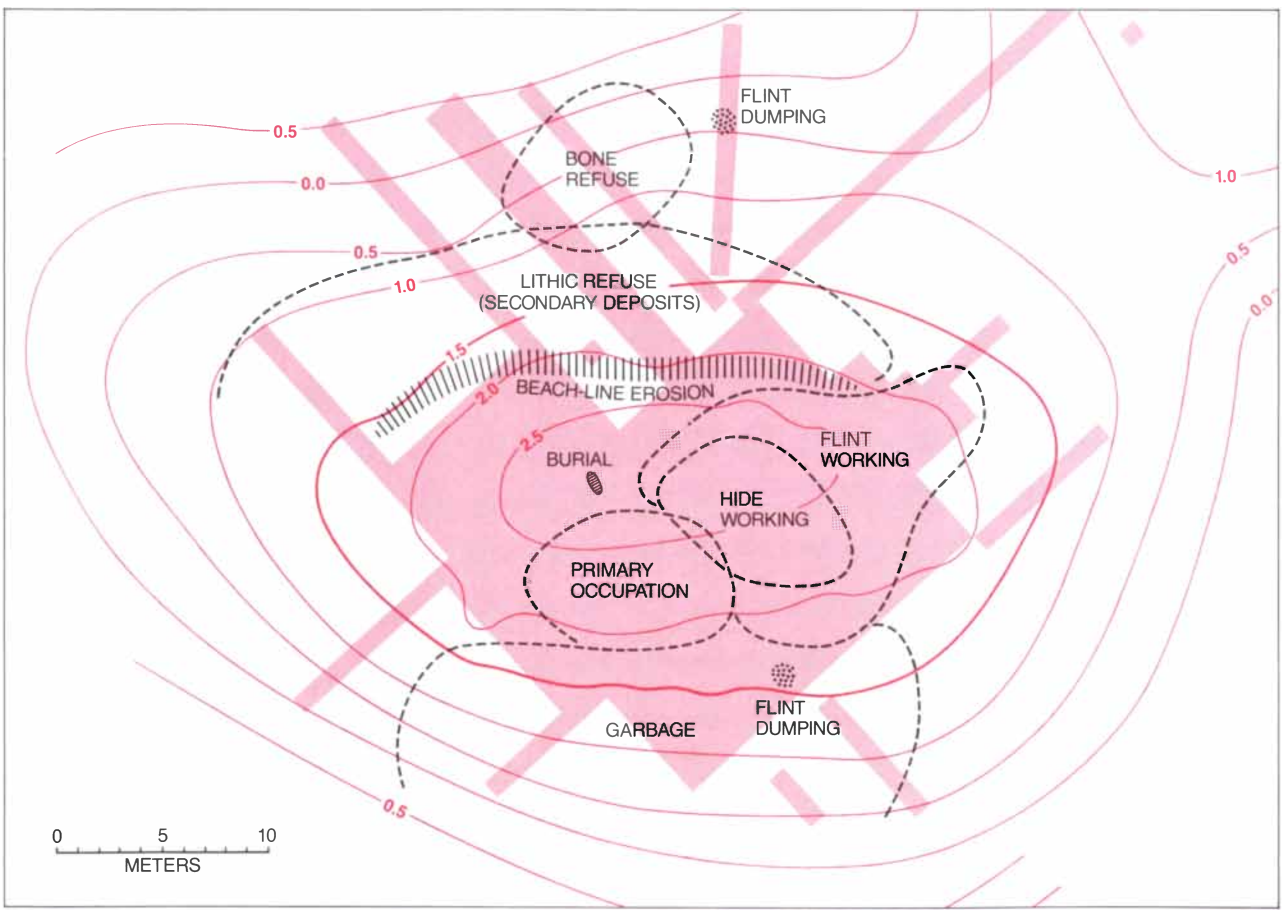
Another major feature of Vaenget Nord is the single burial pit, which we found on the very highest part of the island in front of the primary occupation zone. The pit lacks both a skeleton and the red ochre pigment found in many of the graves in the cemetery excavated in 1975. Nevertheless, it seems probable that the pit was once a grave: its size and contents (a heavy flint blade and two core axes) resembled those of another burial found earlier at a nearby site. Decomposition has removed the organic part of the burial, leaving only the less destructible flint.

The types of artifacts we found at Vaenget Nord, combined with our reconstruction of the activity areas there, enable us to make some tentative generalizations about the overall function of the site. The island was the focus for a set of related activities. Among them were manufacturing and repairing tools and weapons (as shown by the arrowheads) and butchering animals (as shown by bones and the meat polish on some flint blades). When the butchering was finished, animal by-products were utilized for making tools and other equipment. Some of these items were of bone (as is shown by the wear marks on the burins), others were of hide (as the scrapers and blades indicate) and still others were of antler. The presence of core axes offers evidence that there was woodworking at the site, but the proportion of woodworking tools is low compared with the proportion at other Mesolithic settlements.

What type of living arrangement could have given rise to such a pattern of activity? In answering that question one must keep in mind the fact that the nearby coast of the inlet was heavily settled. Therefore it seems plausible to think Vaenget Nord was a temporary encampment. The size of the island suggests that perhaps only five to 10 people lived there at any one time. Moreover, the evidence collected from the site suggests that the island was occupied in only a few episodes, each lasting for part of a year. It is not possible to determine with precision the seasons when the site was occupied. (The inlet itself was certainly occupied throughout the year.) Yet hazelnuts, which were found on the island, ripen in the fall, and the aquatic species found at the site are generally available in spring or early summer.

Thus some of the pieces begin to fall into place. It seems possible that what we have uncovered at Vaenget Nord is a seasonal camp—occupied in the warmer months—where the residents of a foraging community on the shore of the inlet carried out specialized tasks. Among them were the butchering and skinning of animals, messy tasks that might well have been done away from the main settlement. The prey came from both land and sea, with the sea possibly predominating. The island would quite likely have served as a dock and as a base for repairing equipment used in hunting.

The episodes of occupation of Vaenget Nord must have taken place in a fairly limited period. The radiocarbon dates from the island range from about 5200 to 4800 B.C. Some



evidence of dumping. In the eastern part of the island fragments of rock formed a crude paved gangway. On the basis of such evidence and other artifacts the authors have reconstructed the activity areas of the site. The occupation zone includes a primary occupation area where the dwelling stood and areas for hide and flint working. Around the island were areas where refuse was dumped.

of the latest dates were obtained from four large pairs of posts driven into the soil of the occupation zone. The posts may be the remains of frames or platforms associated with fishing or other aquatic activities. It is possible that these structures were not built during the main occupation of the island. The level of the sea continued to rise, and the island was submerged not long after 5000 B.C. Thereafter the island may have been transformed from a living space to a rich fishing ground, as is suggested by the discovery of a partially crushed and decomposed dugout canoe that had been moored there in about 3700 B.C.

Nor did the changes in the environment cease when the island was covered. As we mentioned above, the removal of the Pleistocene ice sheets led to a rebound of the landmasses that had been covered. That uplift is still continuing, albeit at a much slower pace, and it is more pronounced where the ice was thickest: in the upper Baltic Sea between Sweden and Finland. Northern Denmark, which was closest to the center, continues to rise very slowly out of the sea, removing sites that were once on the coast farther inland. As a result Vaenget Nord is again dry land.

What is the overall significance of the investigations at Vaenget Nord? Perhaps the most significant contribution comes from the mapping of the activity areas, one of the most precise such plots available for any Mesolithic site. Because of the limited occupation of the island surface, it was possible to readily observe distinct zones where specific activities took place. Such patterns are generally obscured on settlements with longer and more intense occupations. The island itself was part of a sedentary community onshore, and such communities are examples of the increasing complexity of human society in southern Scandinavia just prior to the introduction of farming. In sum, the work at Vaenget Nord has contributed to a growing appreciation of the complexity of archaeological sites and of the richness of the coastal cultures that flourished in the moment before agriculture arrived in Europe's northern zone.

WOOD POST is one of several made of alder excavated at Vaenget Nord. This one is 30 centimeters in diameter (each division of the bar measures 20 centimeters). Stone axes were employed to sharpen the post to a point so that it could be driven directly into the island's sediments. Such posts, which may have extended two meters or more above the surface of the island, could have formed the framework for platforms or other structures.

THE AMATEUR SCIENTIST

Calculating the distance to the sun by observing the trail of a meteor

by Jearl Walker

Can the distance to the sun be determined without optical instruments or any other modern equipment? Joseph L. Gerver of Rutgers University has devised a method by which a lower limit can be placed on the mean separation between the earth and the sun. He needs no more than paper, pens and a ruler. A star map is convenient but not essential. With these simple materials and much patience Gerver ascertained that the sun must be at least 65 million kilometers away, which is about half the actual mean distance of 150 million kilometers.

Gerver's scheme involves observing a meteor as it penetrates the earth's atmosphere. A meteor, which is debris from a comet or a chunk of material from the asteroid belts, heats up rapidly as it falls through the atmosphere, becoming so hot that its glow is visible from the ground. Nearly all meteors burn up before they leave the upper atmosphere.

Gerver's method is to determine a meteor's speed with respect to the earth by dividing the duration of the glow into the length of the meteor trail. If the meteor is orbiting the sun, the upper limit to its speed with respect to the sun is related to the earth's speed around the sun. By measuring the meteor's speed through the atmosphere you can calculate the earth's speed and then the radius of the earth's orbit of the sun.

Gerver's method is put into practice during a time of meteor showers. To apply the method you should arrange for several observers to be separated from one another by tens of kilometers. Have them record the time and duration of any meteors they sight and also mark the path of the meteors on a star map. The duration of a meteor burn should be timed by a chant such as "One one thousand, two one thousand" and so on to count off the seconds. Later examine the collected data for any common sighting. If you find one, you can employ the relative positions of the observers and their measurements of the meteor to calculate the height of the meteor's end point, which is where it was last seen.

At this stage Gerver introduces a check on the results. From the computed end point and the observers' perspectives of it he calculates the compass headings between the observers. If the calculated headings approximate the true ones, he knows he is on the right track.

You see meteors against the celestial sphere, an enormous imaginary structure centered on the earth. Since the stars are so distant, they seem to lie on the inside surface of the sphere. The point there at which a meteor shower appears to originate is called the radiant. The shower takes its name from the nearest constellation. In the course of a shower you may see meteors streaming in all directions from the radiant. The appearance is an illusion. The trails are in fact approximately parallel. Their apparent divergence arises from your different perspective of the various paths.

One advantage of having several distant observers is that you can extrapolate the observed trails backward until they cross at approximately the radiant. Gerver extrapolates the trails by holding a straightedge against the night sky, aligning it first along the trail recorded by one of the observers and then along the trail recorded by another observer. He extrapolates each trail to the point where they all cross. (On a flat star map the extrapolations are curved lines because of the distortions imposed in mapping the curved surface of the celestial sphere onto a flat surface.) After locating the radiant you can calculate the height of the meteor when it was first seen and the length of its trail.

The meteor's speed in relation to the earth is the ratio of the calculated length of its trail to the duration of its burn. The speed is only approximate because the duration of the burn is estimated. You could greatly improve the precision of the experiment by timing the burn with an instrument. Better yet, you could videotape a meteor shower and then replay it to time a burn. Such refinements, however, would spoil the fun. Gerver's aim is to calculate the distance to the sun without instruments.

The next step of computation requires a knowledge of the orbital characteristics of meteors. Many meteors, certainly those that are associated with repeated showers, move in an elliptical orbit around the sun before the earth's gravity captures them. When one of them passes through the earth's orbit, its speed with respect to the sun can be no more than $\sqrt{2}$ times the earth's speed. A meteor with a greater speed orbits the sun only once and then flies off to an effectively infinite distance from the sun.

Since Gerver limits his observations to meteors associated with repeated showers, he assumes that any meteor figuring in his calculations will be moving at a speed (with respect to the sun) no greater than the upper limit. Armed with this limit and the measured speed of the meteor with respect to the earth, he sets out to calculate the speed of the earth.

The velocity of an object is a vector: a line with an orientation that represents the direction of the object and a length that represents the object's speed. The velocities of the meteor with respect to the earth and to the sun and the earth's velocity in relation to the sun can be represented in what is termed a vector diagram. The length of the earth's velocity vector is arbitrarily drawn. The length of the meteor's velocity with respect to the sun is $\sqrt{2}$ times as much, under the assumption that the speed of the meteor approximates the upper limit.

The third vector, which extends from the head of the earth's velocity vector to the head of the vector representing the velocity of the meteor with respect to the sun, represents the meteor's velocity in relation to the earth. Its size is the speed you calculate from observations of the meteor trail. Note that the vectors form a triangle for which one side is known and the other sides represent the unknown speed of the earth. To find the earth's speed from the triangle, you must first determine trigonometrically the angle between the earth's velocity and the meteor's observed velocity.

The meteor's velocity in relation to the earth points toward you from the radiant. Mark the radiant on a star map. Determining the direction of the earth's velocity by means of the map is somewhat trickier. The map is marked in right ascension along one axis and in declination along the other axis. The declination is measured in degrees. During the night stars on the celestial sphere move through your view parallel to a line of declination. Right ascension is often measured in units of hours because the celestial sphere rotates through a certain angle in one hour. Since a full rotation of 360 degrees takes 24 hours, each hour of rotation is equivalent to 15 degrees.

Celestial objects can be positioned on the map or in the sky by means of declination and right ascension. Imagine an extension of the earth's Equator to the inside surface of the sphere. Also imagine a spherical grid, similar to the latitude and longitude lines on a map, laid on that surface. The grid represents declination (analogous to latitude) and right ascension (analogous to longitude).

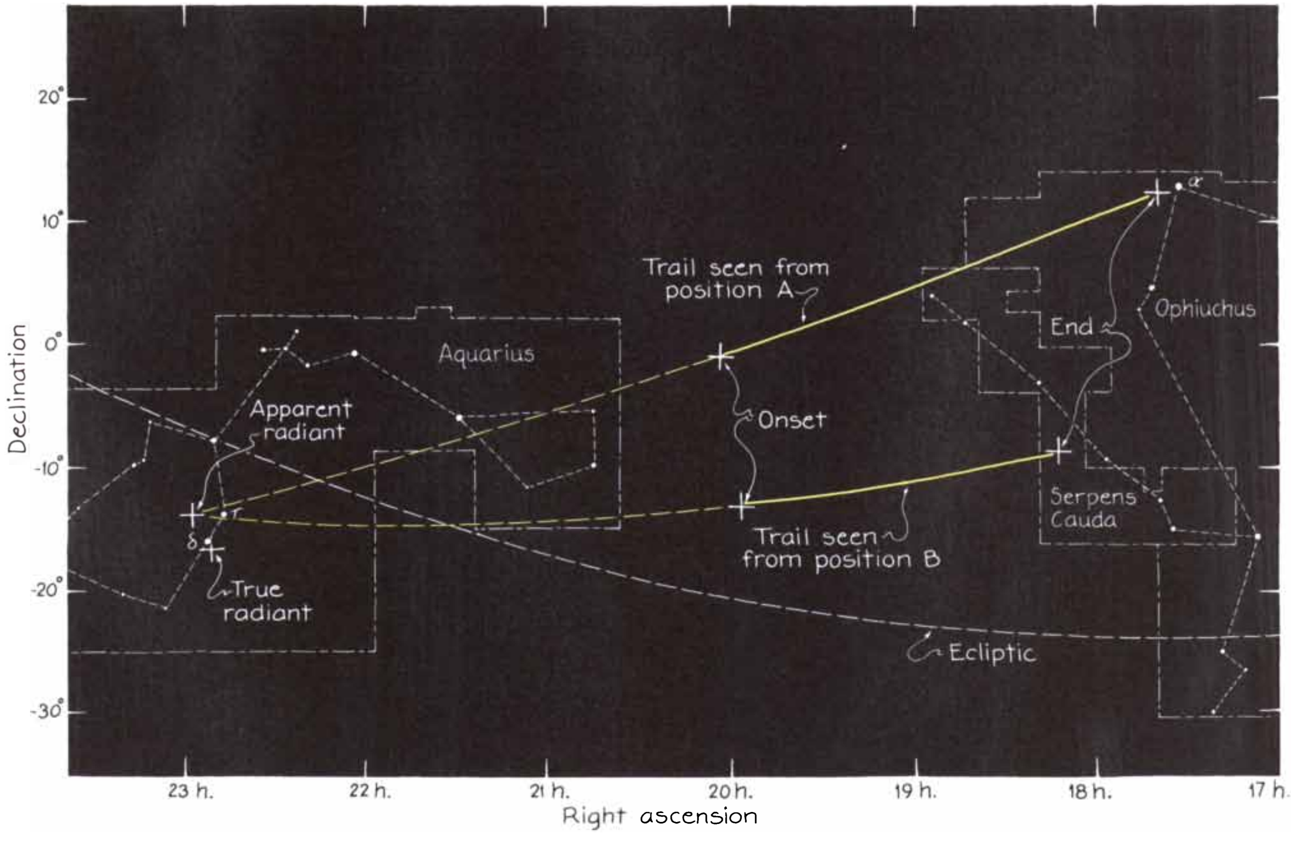
In the course of a year the sun seems to move with respect to the celestial sphere along a curved path called the ecliptic; the movement is measured in degrees. (The motion is, of course, only apparent because the earth in fact orbits the sun in what is called the ecliptic plane.) For the day of your meteor observations figure out where the sun is on the ecliptic. Then move 90 degrees to the east along the ecliptic. It is from the corresponding position in the sky that the earth's velocity vector then points.

You now have determined two velocity vectors on the star map, one for the meteor's velocity with respect to the earth and another for the earth's velocity in relation to the sun. Measure the distance between the two marks in units of degrees according to the scale on either of the axes. For example, you can mark the distance on the edge of a sheet of paper and then align the edge along the declination axis to convert the distance into degrees. The result is only an estimate of the angle between the two vectors because of the distortion of the flat map.

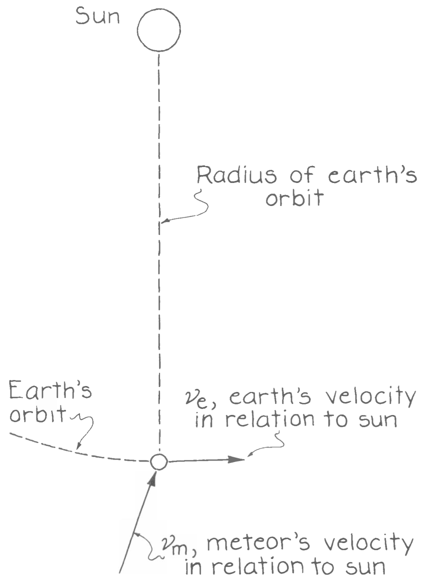
The supplement of the angle you measure is the angle in the vector diagram between the sides representing the earth's speed and the measured meteor speed. Substitute that angle and the lengths of the triangle's sides into the trigonometric cosine rule, which relates the sides to the cosine of one of the angles, to solve for the only remaining unknown, the earth's speed. The calculation yields a lower limit to the earth's speed because it includes the assumption that the meteor's speed with respect to the sun is at its upper limit. The meteor could be slower, in which case the earth's speed would be greater than you had calculated.

Since the earth's orbit is nearly circular, the circumference of the orbit is approximately $2\pi r$, where r is the distance between the sun and the earth. The product of the earth's speed and one year (expressed in suitable units) equals the circumference. From this relation you can calculate a lower limit to the distance to the sun.

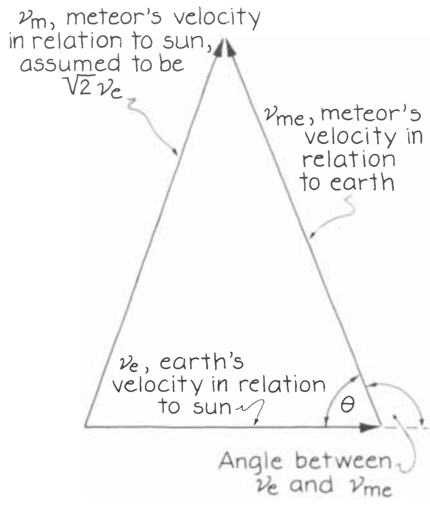
On three occasions over three years Gerver collected data during meteor showers. He was assisted by his brother Michael and Stephen Alessandrini, Marcus Wright, Jonathan Singer, L. Thomas Ramsey, Herb Doughty, Peter Gaposhkin, Ned Phipps, Agnis Kaugars and Cynthia Burnham. He needed such a throng of observers because the chance that widely separated observers would sight the same meteor was fairly small. The group encountered problems such as overcast skies and traffic jams. The first time Gerver attempted the experiment he and an



Meteor trails plotted on a star map



A meteor and the earth's orbit



$$(\sqrt{2}\,v_e)^2 = v_e^2 + v_{me}^2 - 2v_e v_{me} \cos\theta$$

A vector diagram of velocities

assistant 10 miles away spotted a meteor at the same time and in the same part of the sky. When Gerver made his computations, however, the distance to the meteor turned out to be negative. Apparently he and the assistant saw different meteors.

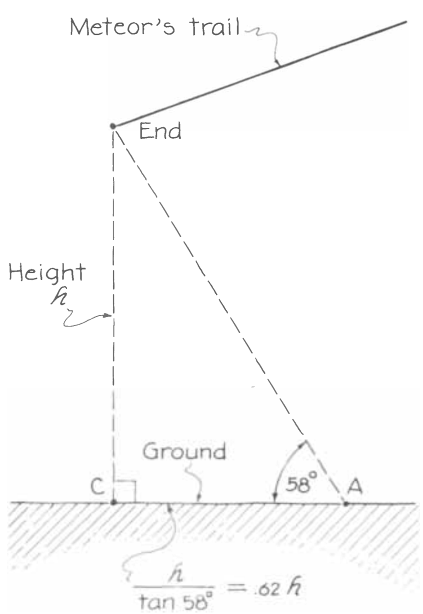
Gerver and his team achieved their first successful observation at 10:56 P.M. Pacific Standard Time last August 11. Gerver, Ramsey and Singer were stationed on Route 17 in the Santa Cruz Mountains midway between Los Gatos and Santa Cruz, Calif., and Michael Gerver was 31 kilometers distant in Palo Alto. Call the first position *A* and the second one *B*. All observers spotted a meteor that was as bright as Jupiter and had an unusually long trail. Although most of the meteors visible that night were part of the Perseid shower, Gerver believes the bright one seen by the observers was part of the Delta Aquarid shower, whose path is perpendicular to that of the Perseid group.

The observers recorded the altitude and azimuth of the meteor's end point. The altitude is the angle between the end point and the point on the horizon directly below it. The azimuth is the horizontal angle measured clockwise (as it would be seen from above) from north to just below the end point. From *A* the altitude of the end point was 58 degrees and the azimuth was 225 degrees. From *B* the altitude was 43 degrees, the azimuth 200 degrees.

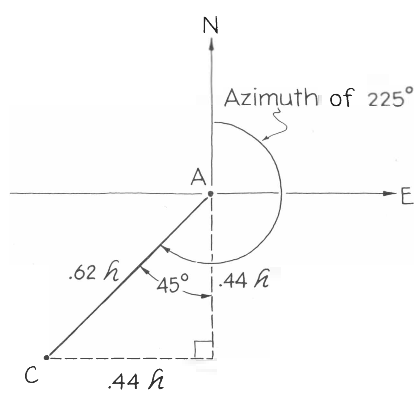
Let *C* be the point on the ground directly below the end point and *h* be the height of the end point. Construct a right triangle with one leg equal to *h* and another leg equal to the distance between *A* and *C*. The altitude is the angle between the hypotenuse and the leg along the ground. By means of the tangent of the angle determine the ground-level leg in terms of *h*. It is .62 *h* in length.

Next take an overhead view of *A* and *C*. Construct a right triangle in which the hypotenuse is the *A*-to-*C* separation of .62 *h*, one leg running east and west and the other leg north and south. Since the azimuth to *C* as it is seen from *A* is 225 degrees, the angle between the hypotenuse and the north-south leg must be 45 degrees. The cosine of 45 degrees equals the ratio of the north-south leg to the hypotenuse. Solve the equation for the north-south leg in terms of *h*: the leg is equal to .44 *h*. Similarly solve for the east-west leg in terms of *h* with the aid of the sine of 45 degrees. The east-west leg is also .44 *h*. Hence *A* is .44 *h* north and .44 *h* east of *C*.

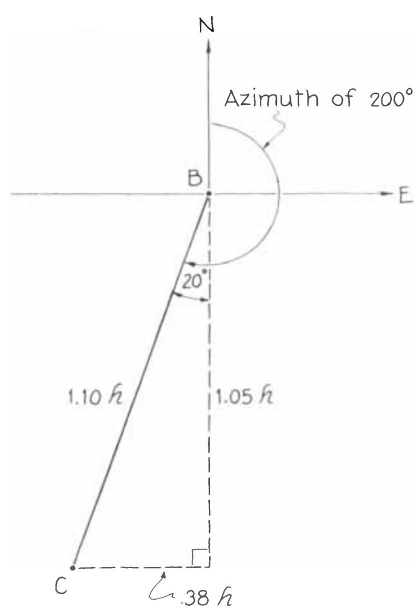
Repeat the procedure for the observations from *B*. The ground separation between *B* and *C* is 1.1 *h*. Since *C* has an azimuth of 200 degrees, the angle in the ground-level triangle between the hypotenuse and the north-south leg is 20 degrees. After solving the appro-



Calculating end-point height



Ground plan for position **A**



Ground plan for position **B**

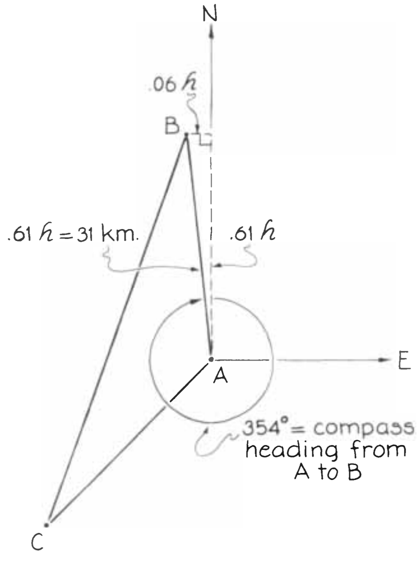
priate trigonometric relations you will find that *B* is 1.05 *h* north and .38 *h* east of *C*.

Next draw a ground-level triangle that includes *A, B* and *C*. Subtract the north-south distance between *A* and *C* from the north-south distance between *B* and *C* to find that *B* is .61 *h* north of *A*. Subtract the east-west distance between *B* and *C* from the east-west distance between *A* and *C* to find that *B* is .06 *h* west of *A*. The north-south and east-west distances between *A* and *B* form the legs of a right triangle that has a hypotenuse equal to the actual distance between the points. By means of the Pythagorean theorem calculate the hypotenuse, which turns out to be .61 *h*. Since this symbolic result must equal the actual separation of 31 kilometers, *h* must be 50 kilometers. You now have the height at which the meteor was last seen.

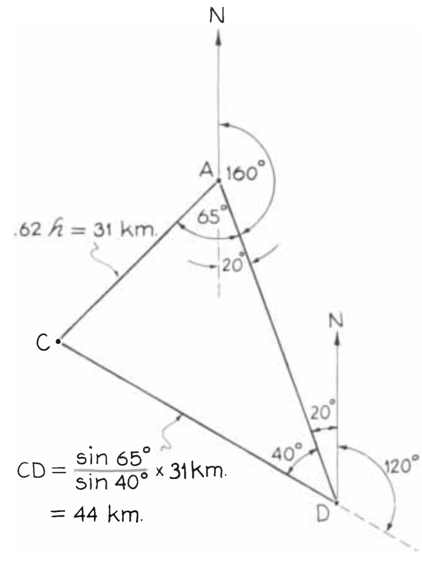
As a check on the calculations, reconsider the ground-level triangle in which the separation of *A* and *B* is the hypotenuse. With *h* equal to 50 kilometers, the north-south leg of the triangle must be 31 kilometers and the east-west leg three kilometers. Since the tangent of the angle between the hypotenuse and the north-south leg equals the ratio of three kilometers to 31 kilometers, the corresponding angle must be six degrees. Hence the compass heading from *A* to *B* must be six degrees less than 360 degrees, or 354. The true compass heading was 345 degrees. Since the discrepancy is only nine degrees, the calculations appear to be trustworthy.

From *A* the beginning of the meteor trail was at an altitude of 52 degrees and an azimuth of 160 degrees. If Gerver were certain that both sets of observers spotted the beginning of the trail simultaneously, he could find that point the same way he did the end point. Without that certainty he must find the starting point by means of the radiant. Several nights after the observations he extrapolated the observed paths of the meteor back over the sky by means of his straightedge. The trails intersected just east of the star Tau Aquarii at an altitude of 20 degrees and an azimuth of 120 degrees. This approximate radiant is a few degrees from the star Delta Aquarii, which Gerver believes is the true radiant for the meteor.

To calculate the length of the trail and the height of the beginning point as Gerver did construct an overhead view of the trail's projection onto the ground. Let *D* be the point on the ground directly below the trail's onset point. The orientation of the trail is set by the azimuth of the radiant, which is 120 degrees. Find the angle *CAD* within the triangle *ACD* by subtracting the azimuth of *D* from the azimuth of *C*. Find the angle *ADC* on the basis that the acute angle between *AD* and the north-south line is 20 degrees.



The plan for **A** *and* **B**



Ground plan for a meteor trail

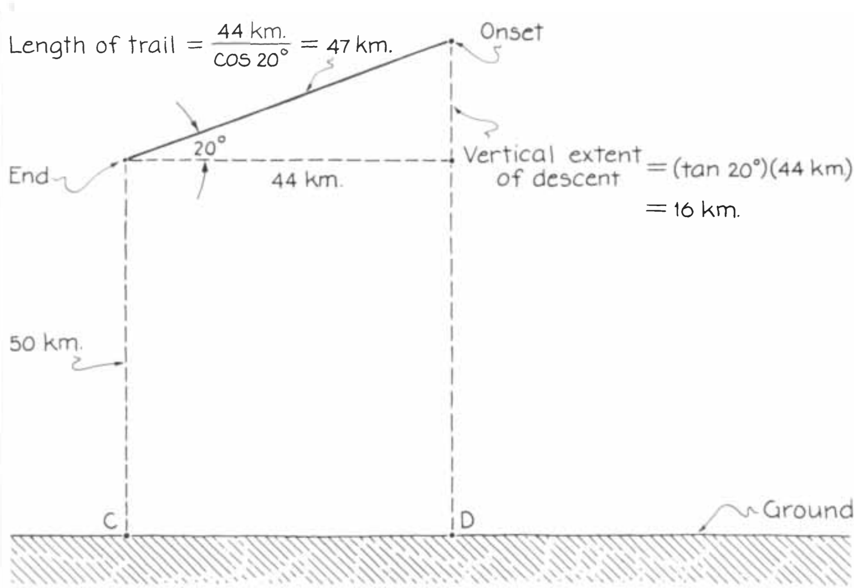
From a previous calculation you know that the length of *AC* is .62 *h*, which is 31 kilometers. By means of the trigonometric law of sines, which relates the sines of two of the angles to the sides of the triangle opposite those angles, find the length of *CD* (44 kilometers). Next draw a right triangle in which the hypotenuse represents the length of the meteor trail, one leg represents the distance between *C* and *D* and the other leg represents the vertical extent of the meteor's trail. The angle between the trail and the horizontal is the altitude of the meteor's radiant (20 degrees). Calculate the trail's length with a cosine function; the answer works out to 47 kilometers. With a tangent function calculate the vertical extent of the trail (16 kilometers). You have now determined the length and orientation of the meteor's trail.

The meteor was visible for about 2.5 seconds. In order to travel 47 kilometers in that time, its speed with respect to the earth must be about 19 kilometers per second. On the night of the observations the earth was moving toward the constellation Aries. Gerver found from a star map that the angle between the earth's velocity in relation to the sun and the meteor's velocity with respect to the earth was about 112 degrees. From a triangle of those two vectors and the calculated speed of 19 kilometers per second you can calculate that the earth's speed with respect to the sun must be at least 13 kilometers per second. Divide the product of the speed and the number of seconds in a year by 2π. The result (65 million kilometers) is a lower limit for the distance to the sun.

Gerver puts forward several possible explanations for the fact that the result is only about half the true distance to the sun. The meteor may not have been traveling at the upper limit of its speed in relation to the sun before it entered the atmosphere. Moreover, the slowing effect of the atmosphere is not considered in the calculation. The most serious source of error is the estimate of the duration of the burn. Gerver believes it may be off by as much as 20 percent.

Gerver also observed Perseid meteors, but he had no distant assistants. Perseid meteors follow the orbit of comet Swift-Tuttle, which has a period of 120 years. When they pass through the earth's orbit, their speed in relation to the sun should be near the upper limit of a sun-orbiting object. By extrapolation Gerver estimated their radiant to be a few degrees east of Epsilon Cassiopeiae, which is about 10 degrees from the accepted radiant near Eta Persei. One of the meteors seemed to move from Eta Herculis to Beta Herculis in about .25 to .5 second. At the time the altitude of Eta Herculis was 69 degrees and the azimuth was 274 degrees; the altitude of Beta Herculis was 58 degrees and the azimuth was 240 degrees. The radiant estimated by Gerver was at an altitude of 25 degrees and an azimuth of 22 degrees.

Assuming that the initial height of the meteor was 66 kilometers, Gerver



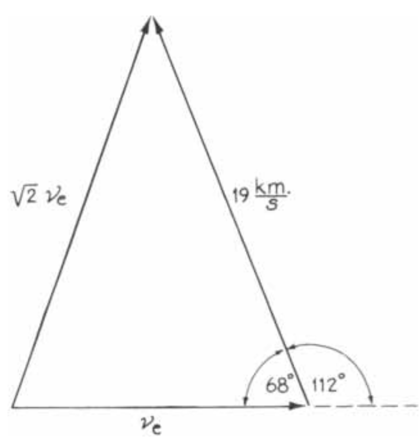
Calculations for the trail

calculated that the final height was 57 kilometers and the total length of the trail was 23 kilometers. Estimating the burn time to be .35 second, he found that the meteor's speed in relation to the earth was 65 kilometers per second. The estimated radiant was about 50 degrees from the direction of travel of the earth. From a vector diagram he calculated a lower limit to the earth's velocity in relation to the sun as being 36 kilometers per second. That result puts the distance to the sun at 180 million kilometers.

Gerver employed star maps, but he points out that you can make altitude measurements with a primitive astrolabe. As a trial he sighted stars with a pad of paper held vertically with one bottom corner in front of his right eye and the bottom of the pad exactly horizontal. He rolled his eye upward until he saw a star aligned with a point on the far vertical edge of the pad. He marked the apparent position of the star on that edge of the paper. With a ruler he then measured the width of the pad and the distance from the bottom edge to the mark. From the ratio of the lengths he calculated the angle between the horizontal and the direction to the star. He found that for any given star he was able to ascertain the angle to within five degrees of the true altitude of the star.

Gerver offers several guides on how best to repeat his experiment. The observers should be familiar with the constellations so that they can accurately locate a meteor trail in the night sky and mark it on a star map. At least two people should be at each observing site. One person should record the duration of the burn while the other memorizes the trail before marking it on the map. Having more than two observing sites improves the accuracy of the calculations.

The weakest stage in Gerver's procedure is estimating the duration of the burn, particularly when it is less than a second. Sometimes, after a meteor had disappeared, Gerver attempted to sweep a finger of his outstretched arm across the sky at about the speed of the meteor. During the sweep he chanted to measure the time. The duration of the meteor burn is the product of the duration of the sweep and the ratio of the angle through which the meteor passed to the angle through which his finger moved.



$$2v_e^2 = v_e^2 + \left(19\frac{km.}{s}\right)^2 - 2v_e\left(19\frac{km.}{s}\right)\cos 68°$$

Earth-velocity vectors


DEXTER'S
HIGH-TECH
ICE-CREAM
frozen yogurt
ICE CREAM BARS
ITALIAN ICES
SNOW CONES
ICE

BIBLIOGRAPHY

Readers interested in further explanation of the subjects covered by the articles in this issue may find the following lists of publications helpful.

COMPUTER RECREATIONS

INTRODUCTION TO NERVOUS SYSTEMS. Theodore Holmes Bullock, Richard Orkand and Alan Grinnell. W. H. Freeman and Company, 1977.

SMALL SYSTEMS OF NEURONS. Eric R. Kandel in *Scientific American,* Vol. 241, No. 3, pages 66–76; September, 1979.

VEHICLES: EXPERIMENTS IN SYNTHETIC PSYCHOLOGY. Valentino Braitenberg. The MIT Press, 1984.

DYSLEXIA

THE PSYCHOLOGY OF READING. Eleanor J. Gibson and Harry Levin. The MIT Press, 1975.

DYSLEXIA: THEORY AND RESEARCH. Frank R. Vellutino. The MIT Press, 1979.

SPEECH, THE ALPHABET AND TEACHING TO READ. I. Y. Liberman and D. Shankweiler in *Theory and Practice of Early Reading: Vol. 2,* edited by Lauren B. Resnick and Phyllis A. Weaver. Lawrence Erlbaum Associates, 1980.

BIOBEHAVIORAL MEASURES OF DYSLEXIA. Edited by David B. Gray and James F. Kavanagh. York Press, 1985.

THE STRUCTURE OF POLIOVIRUS

PHYSICAL PRINCIPLES IN THE CONSTRUCTION OF REGULAR VIRUSES. D. L. D. Caspar in *Cold Spring Harbor Symposia on Quantitative Biology,* Vol. 27, pages 1–24; 1962.

PORTRAITS OF VIRUSES: THE PICORNAVIRUSES. Joseph L. Melnick in *Intervirology,* Vol. 20, Nos. 2–3, pages 61–100; 1983.

PRINCIPLES OF VIRUS STRUCTURE. Stephen C. Harrison in *Virology,* edited by Bernard N. Fields. Raven Press, 1985.

THREE-DIMENSIONAL STRUCTURE OF POLIOVIRUS AT 2.9 Å RESOLUTION. J. M. Hogle, M. Chow and D. J. Filman in *Science,* Vol. 229, No. 4720, pages 1358–1365; September 27, 1985.

COOLING AND TRAPPING ATOMS

LASER-COOLED AND TRAPPED ATOMS. W. D. Phillips. National Bureau of Standards, Special Publication No. 653, 1983.

ATOMIC BEAMS STOPPED BY LASER LIGHT, AND TRAPPED. Bertram M. Schwartzchild in *Physics Today,* Vol. 38, No. 6, pages 17–21; June, 1985.

LASER COOLING AND ELECTROMAGNETIC TRAPPING OF NEUTRAL ATOMS. William D. Phillips, John V. Prodan and Harold J. Metcalf in *Journal of the Optical Society of America B,* Vol. 2, No. 11, pages 1751–1767; November, 1985.

LASER EXTREMES PROBE ATOMS AND MOLECULES. Arthur L. Robinson in *Science,* Vol. 230, No. 4727, page 797; November 15, 1985.

SODIUM ATOMS STOPPED AND CONFINED. Arthur L. Robinson in *Science,* Vol. 229, No. 4708, pages 39–41; July 5, 1986.

MONOCULTURE

NO-TILLAGE AGRICULTURE PRINCIPLES. Edited by Ronald E. Phillips and Shirley H. Phillips. Van Nostrand Reinhold Co., 1984.

ORGANIC FARMING: CURRENT TECHNOLOGY AND ITS ROLE IN A SUSTAINABLE AGRICULTURE. Edited by D. F. Bezdicek. American Society of Agronomy, 1984.

SOIL EROSION AND CROP PRODUCTIVITY. Edited by R. F. Follett and B. A. Stewart. American Society of Agronomy, 1985.

OPTICAL NEURAL COMPUTERS

SPECIAL ISSUE ON OPTICAL COMPUTING. In *Proceedings of the IEEE,* Vol. 72, No. 7; July, 1984.

THE COMPLEXITY OF INFORMATION EXTRACTION. Yaser Abu-Mostafa in *IEEE Transactions on Information Theory,* Vol. IT-32, No. 4, pages 513–525; July, 1986.

OPTICAL COMPUTING: A FIELD IN FLUX. Trudy E. Bell in *IEEE Spectrum,* Vol. 23, No. 8, pages 34–57; August, 1986.

COMPLEXITY IN INFORMATION THEORY. Edited by Y. Abu-Mostafa. Springer-Verlag, 1987.

THE RIFTING OF CONTINENTS

A MODEL FOR DEVELOPMENT OF RED SEA. James R. Cochran in *American Association of Petroleum Geologists Bulletin,* Vol. 67, No. 1, pages 41–69; January, 1983.

PERIDOTITE COMPOSITION FROM THE NORTH ATLANTIC: REGIONAL AND TECTONIC VARIATIONS AND IMPLICATIONS FOR PARTIAL MELTING. Peter J. Michael and Enrico Bonatti in *Earth and Planetary Science Letters,* Vol. 73, No. 1, pages 91–104; April, 1985.

PUNCTIFORM INITIATION OF SEAFLOOR SPREADING IN THE RED SEA DURING TRANSITION FROM A CONTINENTAL TO AN OCEANIC RIFT. E. Bonatti in *Nature,* Vol. 316, No. 6023, pages 33–37; July 4, 1985.

PERIDOTITES FROM THE ISLAND OF ZABARGAD (ST. JOHN), RED SEA: PETROLOGY AND GEOCHEMISTRY. Enrico Bonatti, Giulio Ottonello and Paul R. Hamlyn in *Journal of Geophysical Research,* Vol. 91, No. B1, pages 599–631; January 10, 1986.

THERMOREGULATION IN WINTER MOTHS

TEMPERATURE REGULATION OF THE SPHINX MOTH, *MANDUCA SEXTA:* II, REGULATION OF HEAT LOSS BY CONTROL OF BLOOD CIRCULATION. Bernd Heinrich in *The Journal of Experimental Biology,* Vol. 54, No. 1, pages 153–166; February, 1971.

MORPHOMETRICS, CONDUCTANCE, THORACIC TEMPERATURE, AND FLIGHT ENERGETICS OF NOCTUID AND GEOMETRID MOTHS. Timothy M. Casey and Barbara A. Joos in *Physiological Zoology,* Vol. 56, No. 2, pages 160–173; April, 1983.

FLIGHT OF WINTER MOTHS NEAR 0°C. Bernd Heinrich and Thomas P. Mommsen in *Science,* Vol. 228, No. 4696, pages 177–179; April 12, 1985.

THERMOREGULATION BY WINTER-FLYING ENDOTHERMIC MOTHS. Bernd Heinrich in *The Journal of Experimental Biology,* in press.

A MESOLITHIC CAMP IN DENMARK

THE EARLY POSTGLACIAL SETTLEMENT OF NORTHERN EUROPE: AN ECOLOGICAL PERSPECTIVE. Edited by Paul Mellars. Duckworth, Ltd., 1978.

MESOLITHIC PRELUDE: THE PALEOLITHIC-NEOLITHIC TRANSITION IN OLD WORLD PREHISTORY. Grahame Clark. Edinburgh University Press, 1980.

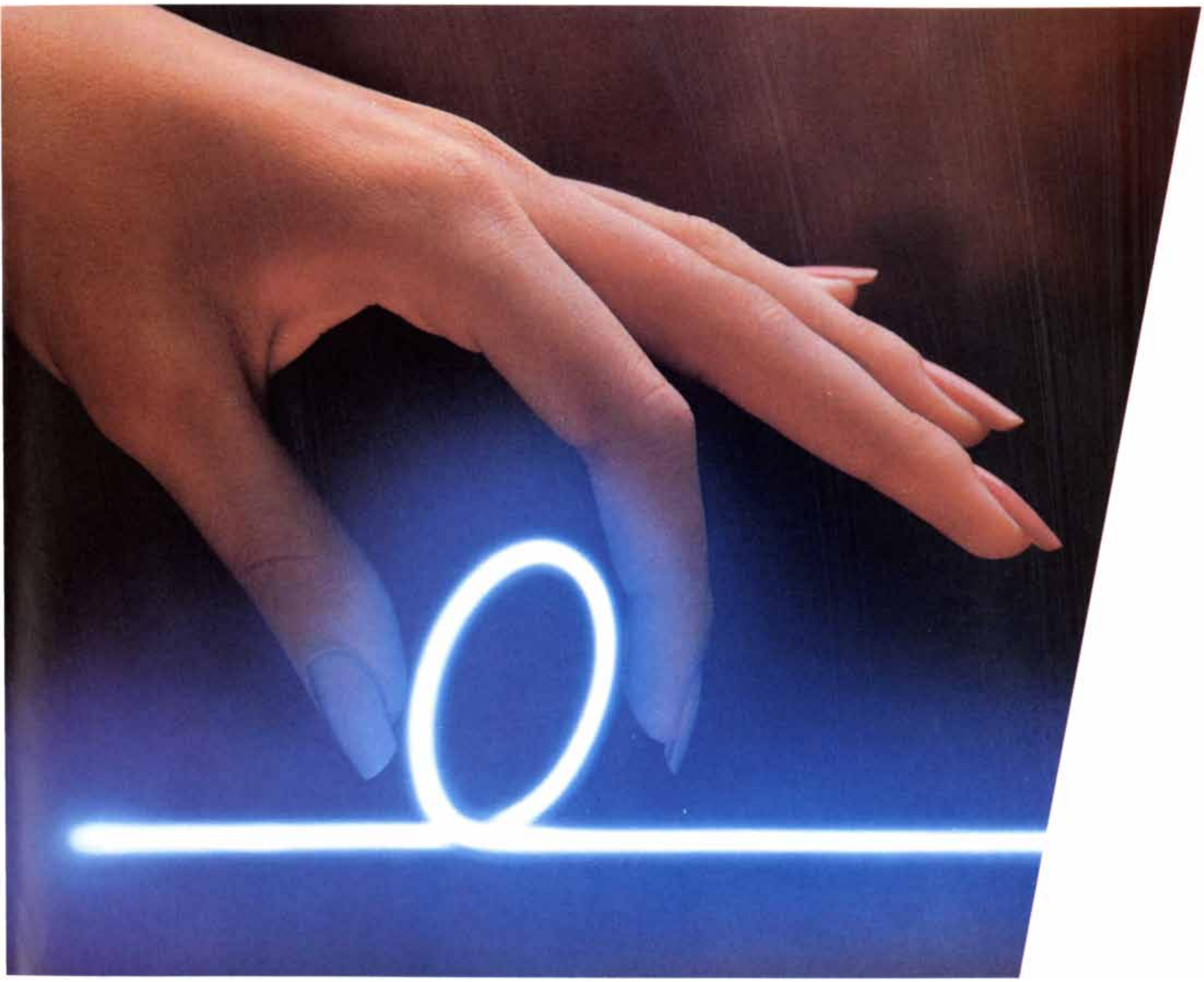
THE EUROPEAN MESOLITHIC. T. Douglas Price in *American Antiquity,* Vol. 48, No. 4, pages 761–778; October, 1983.

THE AMATEUR SCIENTIST

BASIC PHYSICS OF THE SOLAR SYSTEM. V. M. Blanco and S. W. McCuskey. Addison-Wesley Publishing Co., 1961.

THE AMATEUR SCIENTIST. Jearl Walker in *Scientific American,* Vol. 240, No. 5, pages 172–182; May, 1979.

THE COMPANIES OF AMERITECH BRING YOU A NETWORK SMART ENOUGH TO RESPOND TO YOUR NEEDS.



The companies of Ameritech are helping you take greater control over your telecommunications services.

Working with you, we can customize our fiber optic and digital network to meet your specific needs.

We can help you choose from a growing range of digital services that transmit voice and data with amazing speed, volume and accuracy. You'll command a network with the intelligence to give you information about its use and cost. Smart enough to give you what are virtually private digital lines at the lower cost of shared lines.

Talk to our Ameritech Bell companies of Illinois, Indiana, Michigan, Ohio and Wisconsin. They have the network technologies, the combined resources and expert people close at hand to provide telecommunications systems you can manage and rearrange as needed.

To put our network at your command, contact any of our Ameritech Bell companies. Or phone us: 1 800 562-2444.

AMERITECH
Helping you communicate℠

The companies of Ameritech: Illinois Bell · Indiana Bell · Michigan Bell · Ohio Bell · Wisconsin Bell · Ameritech Communications
Ameritech Mobile Communications · ADR · Ameritech Services · Ameritech Credit · Ameritech Development · Ameritech Publishing


THE ULTIMATE VISION:
EYES FOR THE BACK OF YOUR HEAD.
To see some accidents coming, you'd have to be born with radar. But you're not, so GM is working to give you the next best thing. Radar for your car. The option to "see" all around you, all at once. It's called Obstacle Detection. Just one of the many safety features GM is developing for cars of the future.
Obstacle Detection will operate by sensors strategically placed all around your car which will sense and warn you of obstacles you're approaching, giving you a chance to avoid them.
At GM, we can't prevent all accidents from happening, but we want to help you steer clear of as many as we can.
THE GM ODYSSEY:
SCIENCE NOT FICTION
GM
SPEED
FRONT
WATCH
SIDE
UNLEADED
FUEL ONLY
OIL PRESSURE