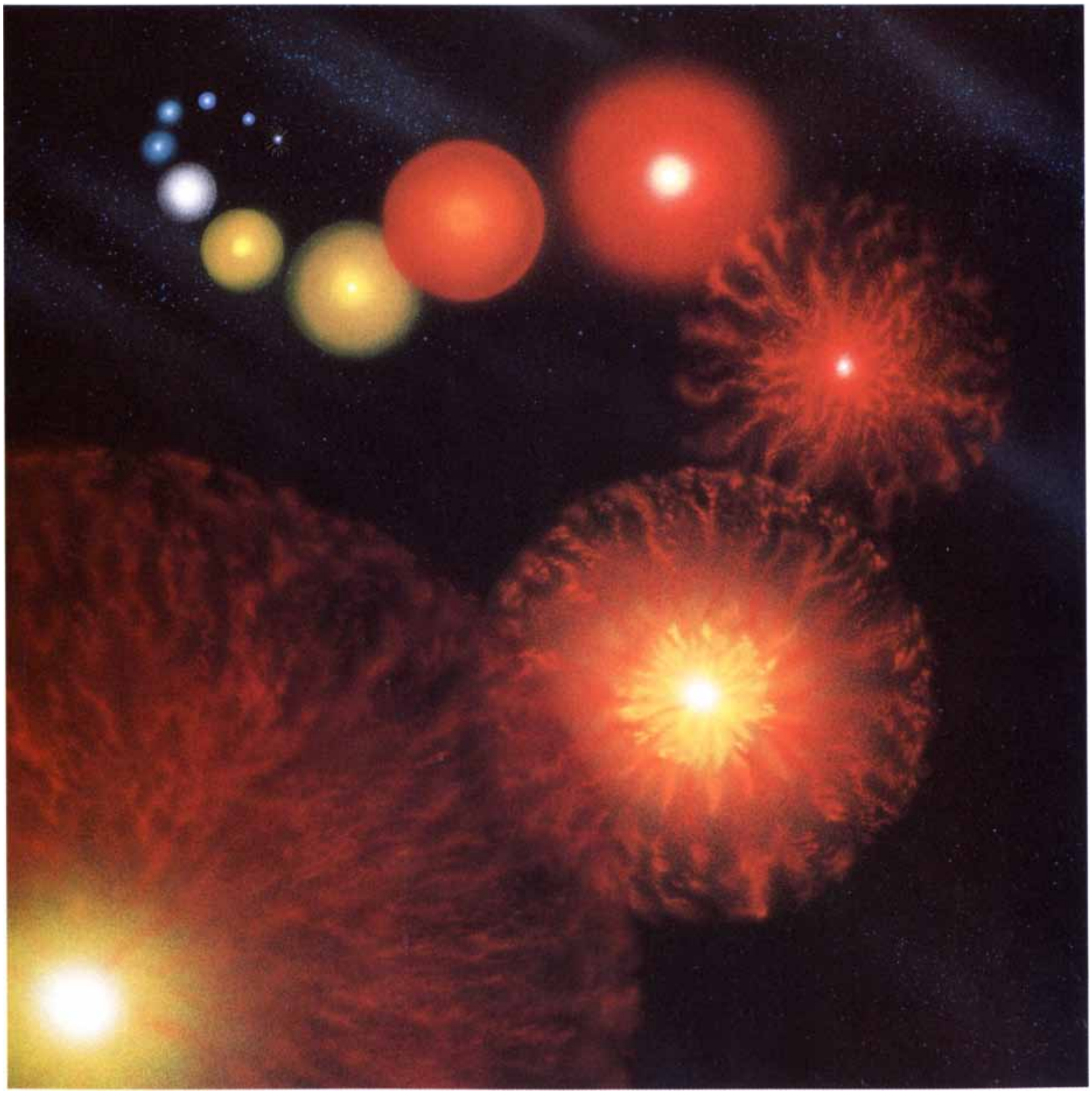
SCIENTIFIC AMERICAN

AUGUST 1989
$2.95

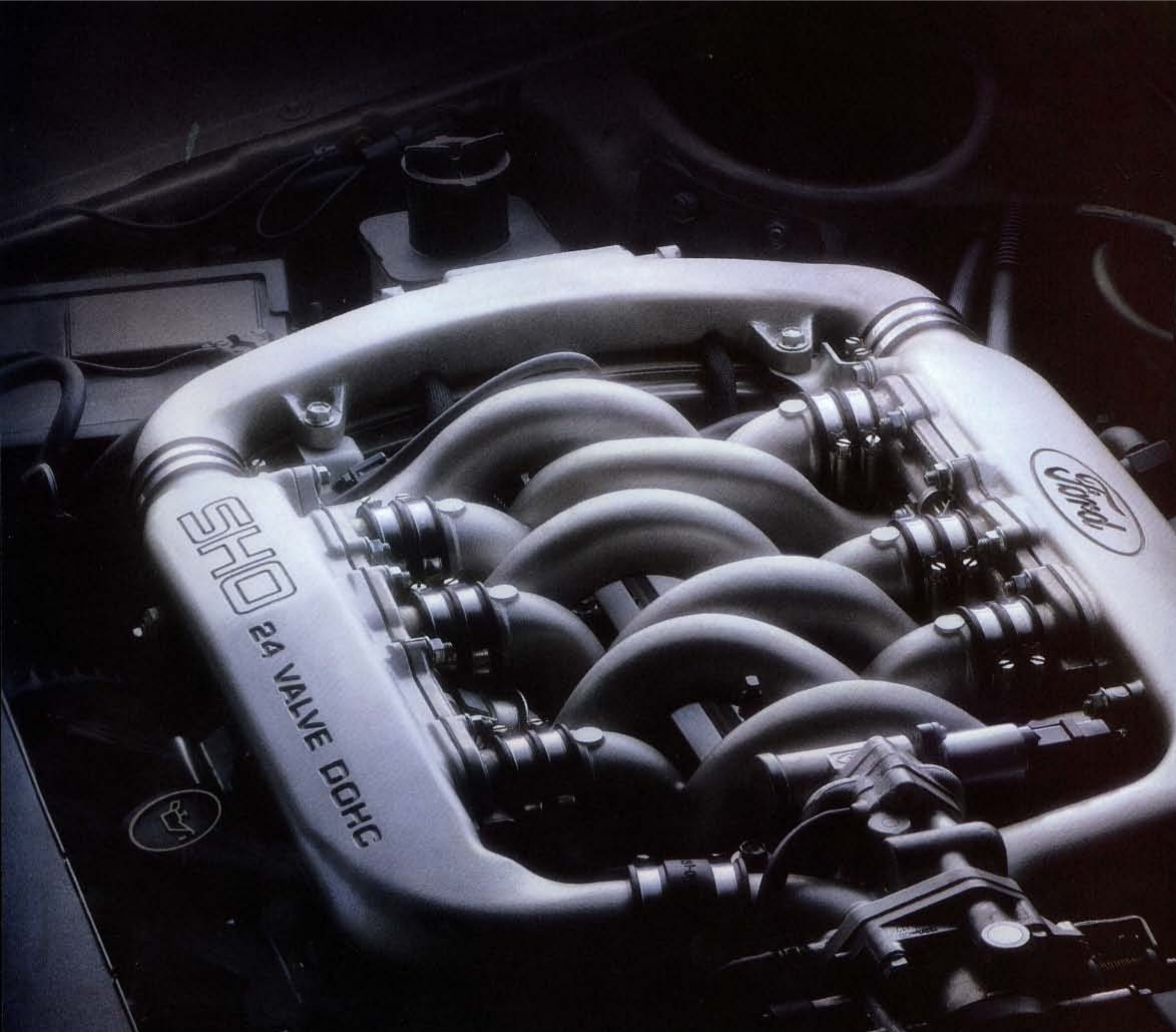
Parallel programming: seven-league boots for the mind.

Spare that rabbit—cell tests for gauging toxic risks.

Cracking Maya writing: new window on an ancient culture.



Supernova life cycle: *new understanding, new challenges from a star exploding in our cosmic neighborhood.*


SHO
24 VALVE DOHC
Ford

# of a four-door sedan inside out.

to control it precisely. So we gave Taurus SHO a 5-speed manual transmission and four-wheel independent sport suspension, for agile handling on demanding roads. And four-wheel disc brakes for stops along the way.

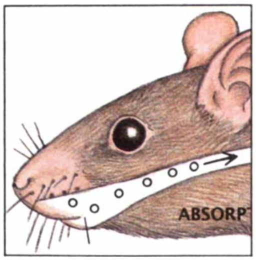
That same precision is reflected in the interior of Taurus SHO. Designed to reduce distractions. Designed to increase your enjoyment of the road. Standard features include a 6-way power adjustable driver's seat and an 8000 RPM tachometer.

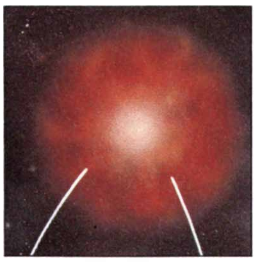
The new limited production Ford Taurus SHO. Conceived to turn your preconceived notions about four-door sedans inside out.

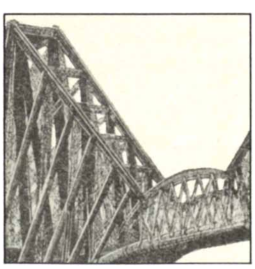
Have you driven a Ford...lately?

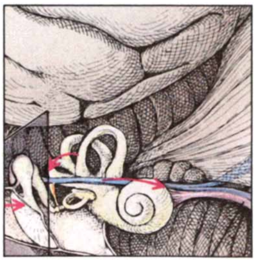
SCIENTIFIC AMERICAN

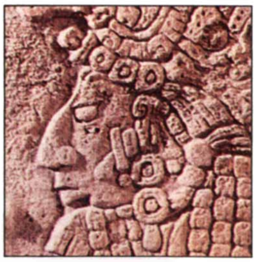
August 1989 Volume 261 Number 2



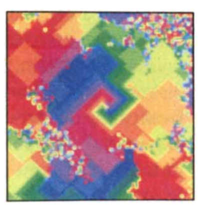








DEPARTMENTS



THE LIFESTYLE RESOURCE™

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

ZONE OF CALM

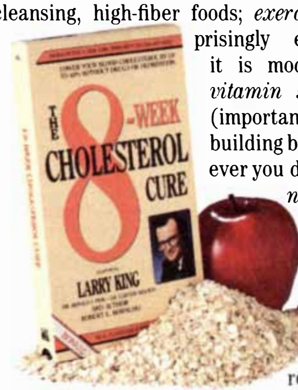
Noise produces stress in the human organism. Today the volume of civilization nearly everywhere seriously interferes with our abilities to relax, read, sleep, concentrate or otherwise function at optimum efficiency. But you needn't be victimized by noise pollution. The *new* Marsona Sound Conditioner electrically synthesizes a variety of pleasing natural sounds that help mask out and reduce the annoyance of unwanted noise. A 5 inch speaker brings you the sounds of ocean surf, summer rain, mountain waterfalls — as close to nature as you will find. You control the volume, the wave pattern, wave or rain rhythm and the seeming nearness or distance of the source. Solid state electronics designed for continuous use. UL listed. Manufacturer's 1 year limited warranty. **$139.95** #2202.

WORRY-FREE WATERING

Water grows things best when it arrives at regular intervals, a fact that sometimes gets by Mother Nature. Now give your lawn or garden the advantage of timely watering. RainMatic 2000 automatically turns water on and off, up to four times a day. Set any length watering interval with your own schedule or use the simple built-in program. Manually override anytime. Great for sprinklers and trickle systems. An absolute necessity during restricted watering. Operates all season on 4-C batteries (not included). Weather-protected. LCD readout with digital clock. Locks on faucet. Manufacturer's 1-yr. warranty. **$49.95** #1081.

DON'T FLIRT WITH A HEART ATTACK OR STROKE!

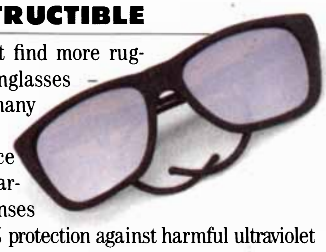
Avoid the danger zone of heart attack and stroke – reduce your deadly cholesterol count. "The 8-Week Cholesterol Cure" is a proven, sensible way: a fascinating 66-minute video based on the New York Times No.1 non-fiction best-seller. It shows you how to benefit from three solid risk-reduction techniques: *better nutrition* (including essential, artery-cleansing, high-fiber foods; *exercise* (as surprisingly effective as it is moderate); and *vitamin supplements* (important defensive building blocks). Whatever you do, get a plan *now* to reduce risk. Over one million Americans have started this doctor recommended program. **$29.95** #2610.



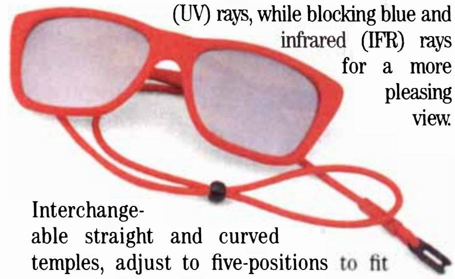
FROM CHINA TO YOUR HEALTH

Ancient mandarins dating back 800 years believed that these Chinese Exercise Balls induced well-being of the body and serenity of spirit. These treasured gifts were given to President Reagan and his wife while visiting the Peoples Republic of China. The Chinese say that rotating the balls in the palm of each hand stimulates the fingers and acupuncture points, and improves circulation of vital energy throughout the body. Sports enthusiasts, musicians, computer users and health-conscious people everywhere consider them great muscle conditioners. Arthritis sufferers feel a decided benefit from this gentle but challenging exercise. Very effective for relaxation and meditation, Chinese Exercise Balls emit a distantly mysterious chime as you turn them. Beautifully handcrafted, 45mm. hollow polished chrome balls are perfectly weighted and fit comfortably into the average man's or woman's hand. In exquisite silk brocade box. **$29.95** #1701.

SOLARMAXX™...VIRTUALLY INDESTRUCTIBLE



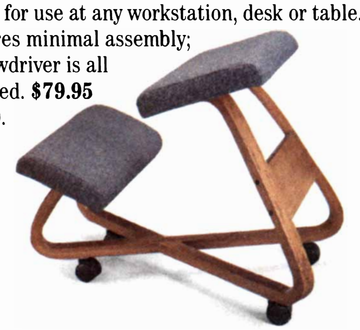
You won't find more rugged sunglasses – with this many advantages – in any price range. Solarmaxx™ lenses provide 100% protection against harmful ultraviolet (UV) rays, while blocking blue and infrared (IFR) rays for a more pleasing view.



Interchangeable straight and curved temples, adjust to five-positions to fit you precisely for maximum comfort and no slippage. The matte-finished, hard rubber frames won't break if they are twisted, sat on, stepped on, jumped on, or even closed in the car door. The neck cord adjusts – and can be attached so it won't slip off the earpieces. Smartly styled in your choice of black or red. Deluxe drawstring case and cleansing cloth included. **$29.95** Black #2580; Red #2590.

SIT UP STRAIGHT!

Sitting in a conventional chair, bent over a desk or table for long periods of time can stress your spine and back muscles. Posture or "back" chairs have long been accepted as a way to relieve your back from the pain of cramping or kinking up due to this stress. Our posture chair is designed to allow your body to properly align itself by shifting weight from your lower back to your knees, making long work periods far more comfortable. *The Lifestyle Resource* is one of the first to introduce the Criss-Cross Posture Chair. This new and innovative design with a criss-crossed base, provides added stability and an easy exit from your seat. The chair is crafted of laminated hardwood with an attractive oak finish. The seat and kneepad are thickly padded and fully upholstered in a classic neutral grey fabric. The swiveling, carpet-friendly, double wheel casters make maneuvering a breeze. Suited for use at any workstation, desk or table. Requires minimal assembly; a screwdriver is all you need. **$79.95** #2260.



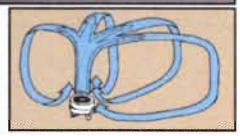
THE LIFESTYLE RESOURCE™

WHICH SPORTS WATCH?

Choose our precision, chrono-style, competition watch by Casio which has everything except a steep, chrono-style price (often from $150 up in pricier models). It's a faithful, sporty companion, strong enough to endure the hard knocks of running, swimming, cycling or washing your car. Precision quartz movement is accurate to ±15 seconds/month. Stainless steel case and water resistant up to 50 meters. Bezel rotates for elapsed timing – easily set with gloves on. Three luminous hands; date window; adjustable PVC wristband may be trimmed to fit. This watch accepts the challenge to combine quality and price – and wins! Manufacturer's 1 yr. limited warranty. Yellow with black #2630; Black with red #2640; Black with blue #2650. Only **$49.95** each.

AERODYNAMIC COOLING

Turbo-Aire's jetstream



Hassock position–venturi effect

The Turbo-Aire fan moves more air and provides more cooling than ordinary fans - you get more than twice the air delivered by oscillating room fans. Patented and computer designed to maximize efficiency and minimize noise and vibration. This fan creates an exceptionally strong, smooth jetstream column that sets a room stirring with secondary air currents. Refreshes better than the intermittent blast of air from an oscillating fan. Aerodynamic housing increases blade-tip efficiency over conventional fans. Floor, table or wall mount, adjusts to any angle. Set in the hassock position, the five bladed, 12″ fan redistributes air in an entire room, making it as useful in winter as in summer. Can reduce air conditioning costs and is 300% more energy efficient than ordinary fans with the same type and size motor. With a 3 speed fan control, the Turbo-Aire is the most advanced 12″ portable house fan. Weighs less than a case of cold soft drinks - and more easily moved! Manufacturer's 2-year limited warranty. UL listed. **$79.95** #2161.

38,000 HOUR LIGHTBULB

Our superior Power Mizers use an exclusive microcontrol switch to extend bulb life *nearly 40 times longer* with virtually no loss of brightness. You'll get 38,000 continuous hours of brilliant light *along with a 24% saving in energy consumption from a common* 100-watt, 1-way bulb rated for 1,000 hours! A blessing in hard to reach fixtures: no need to change bulbs often. Paying for themselves many times over by saving power and bulb costs. UL listed. Manufacturer's 8-yr. warranty. Set of 6 **$19.95** #2001; 12 for **$29.95** #1991.

MOUNTAIN TOP FRESHNESS

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

FUZZ FREE FM RECEPTION

How do you improve FM reception — without the engineering PhD most antennas seem to require? The Terk Pi FM Antenna is the answer for any receiver. Works two ways: on its side it provides excellent low-noise reception throughout the FM range. But to zero in on marginal channels, just stand it up and rotate. Suddenly they're strong and fuzz-free with amazing clarity and depth. The specially designed Schotz circuit boosts signal strength up to 10 times (30 db of gain) without amplifying noisy static. Tested by Audio magazine, it substantially out-performed the standard dipole on nearly 90% of FM channels tried. Simple, unobtrusive; only 5½″ diameter; installs with just a screwdriver. Essential in apartments where exterior antennas are prohibited. "Truly amazing," says Audio. We know you'll agree! UL listed. Manufacturer's limited lifetime warranty. **$79.95** #2550.

THE LIFESTYLE RESOURCE™

FOR FASTEST SERVICE — CREDIT CARD ORDERS CALL TOLL-FREE 24 HRS A DAY **800-872-5200**

DEPT. SFAH09; 921 EASTWIND DR. SUITE 114: WESTERVILLE, OH 43081

SEND TO (PLEASE PRINT)

ADDRESS

CITY STATE ZIP

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

ACCT. # EXP. DATE

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

Shipping Charge covers UPS, handling and insurance for guaranteed delivery. UPS Second Day available for an additional $7.50 per order.

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

Canadian residents please call (614) 794-2662 for ordering information.

UPS 2ND DAY AIR

ORDER WITH CONFIDENCE

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

THE MEDIA DEVELOPMENT GROUP

©1989 C68 ADVERTISEMENT

THE COVER painting follows the evolution of supernova 1987A, the brightest in 383 years (see "The Great Supernova of 1987," by Stan Woosley and Tom Weaver, page 32). The event is shown from the instant the core of the blue progenitor star imploded, generating a pulse of energetic particles and a shock wave that destroyed the star. Over several weeks the shock-heated material expanded and cooled to red heat; within months it had thinned into a vast, tenuous cloud, lighted by radioactive decay.

THE ILLUSTRATIONS

Cover painting by George V. Kelvin

Page	Source
25	S. Varnedoe
26	Patricia J. Wynne
27	Clonetics Corporation (*top left*), Joseph Leighton, Medical College of Pennsylvania (*top right*), S. Varnedoe (*bottom*)
28-29	Patricia J. Wynne
33	David F. Malin, Anglo-Australian Observatory
35	John Maduell, Lawrence Livermore National Laboratory
36-38	George V. Kelvin
39	Arthur Franceschi, Lawrence Livermore National Laboratory; George V. Kelvin (*inset*)
40	George V. Kelvin
43	UPI/Bettmann Newsphotos
44	Tom Prentiss (*top*), Michael B. A. Oldstone (*bottom*)
45-46	Tom Prentiss
47	Michael B. A. Oldstone
48	Murray Greenfield
51-59	David H. Funk
67	The Bettmann Archive
68-72	Andrew Christie
75	Carol Donner
76	S. Allen Counter (*top*), Gabor Kiss (*bottom*)
77	Gabor Kiss
78	Cherly A. Swindoll (*top*), Berit Engström and Agneta Viberg, Karolinska Institute (*bottom*)
79-80	Carol Donner
83	Museo Nacional de Antropología, Mexico City
84-87	Michael Goodman
88	George Stuart
89	Michael Goodman
91	Director of the British Geological Survey
92	Johnny Johnson
93	S. P. Thompson (*top left*), Otto Hahn (*top right*), Yale University Library (*bottom left*), Mrs. Arthur Holmes (*bottom right*)
94	Quesada/Burke (*left*), Bruce Coleman Inc. (*right*)
99-100D	Michael Goodman
102	David Griffeath, University of Wisconsin at Madison
103	Edward Bell (*top*), David Griffeath (*bottom*)
104	David Griffeath

Scientific American (ISSN 0036-8733), published monthly by Scientific American, Inc., 415 Madison Avenue, New York, N.Y. 10017. Second-class postage paid at New York, N.Y., and at additional mailing offices. Authorized as second-class mail by the Post Office Department, Ottawa, Canada, and for payment of postage in cash. Subscription rates: one year $27, two years $48, three years $66 (outside U.S. and possessions add $11 per year for postage). Subscription inquiries: U.S. only 800-333-1199; other 515-247-7631. Postmaster: Send address changes to Scientific American, Box 3187, Harlan, Iowa 51593.

SCIENTIFIC AMERICAN

Established 1845


© 1989 British Airways
London – New York
LUNCH
Selection of canapés
Appetizers
Golden Asetra caviar presented in a wholemeal tartlet
Accompanied by new season Scottish smoked salmon with
a natural filling of curd cheese, cucumber and mint and a
papillote of trout and salmon mousse
Chilled mango with a combination of citrus fruits
Hot hors d'oeuvre featuring a croustade of wild morels,
chanterelles and girolle mushrooms
Soup
Chicken consommé garnished with truffled galantine
Chilled cream of mulligatawny soup
Main courses
Roast fillet of prime beef. Served with Bearnaise sauce
Poached baby lobster prepared with shallots, chablis
and dry vermouth, flavoured with essence of Welsh mussels
and fresh mustard seeds
Breast of duckling with truffled stuffing
Garnished with walnuts and grapes
Vegetarian speciality featuring a timbale of aubergines,
pearl barley and hazelnuts with fresh tomato and basil
Vegetables
Fresh leaf spinach, spring carrots, braised celery,
jetée promenade potatoes and wild rice pilaf
Salad
Mixed seasonal salad. Served with a choice of mustard vinaigrette
or passion fruit and almond dressing
Desserts
Miniature flan of mirabelles glazed with fresh apricot
Served with double cream
Combination of old fashioned vanilla and mocca ice cream
Fruit
Basket of fresh seasonal fruits
Cheese
A fine selection of English cheeses including Stilton,
mild Single Gloucester and goat's cheese. We also recommend
Rougette, a new soft country cheese from Germany
and Chavannes from France
Coffee - Tea
Coffee, decaffeinated coffee or tea. Served with Cape gooseberries,
chocolate truffles and pralines

LETTERS

To the Editors:

I was sufficiently intrigued by Antoni Akahito's plans for a homemade Planck-mass accelerator ["The Amateur Scientist," April] to rush into my basement and start building one. But then I began to worry about something Akahito failed to consider—namely, what to do with all the energy consumed by the accelerator. Akahito's plans show the accelerator itself built on the earth, with energy supplied from orbiting solar satellites. Even a Planck-mass experimentalist cannot escape the second law of thermodynamics, and so the accelerator energy will ultimately manifest itself as heat. A simple calculation shows that the consequences for the earth's climate dwarf anything we can expect from the greenhouse effect.

Akahito's plans require an accelerator power of about a thousandth of the solar luminosity, or 4×10^{23} watts. To maintain a constant temperature, the earth must radiate energy into space at this same rate. Now, the rate at which an object radiates energy increases with temperature, according to the Stefan-Boltzmann law, which says that the radiation rate is proportional to the fourth power of the temperature. The present temperature of the earth is essentially set by its ability to radiate the energy it now receives from the sun; the effects of greenhouse gases, varying surface compositions and other factors are minor but important variations on the basic energy-balance equation. The problem with the Planck-mass accelerator is that its power requirement is *more than two million times* the rate at which the earth now receives energy from the sun.

If the entire surface of the earth participates in radiating the accelerator energy, the Stefan-Boltzmann law requires that the fourth power of the earth's absolute temperature increase by a factor of more than two million. Thus, the temperature itself must go up by the fourth root of two million, or a factor of nearly 40. The earth's absolute temperature is now about 300 degrees Kelvin, and so it would have to rise to about 12,000 degrees K—twice the surface temperature of the sun! (The curious reader might want to contemplate how a temperature higher than the sun's could be consistent with the second law of thermodynamics.) If the radiating region is smaller than the entire planetary surface, then that area will be even hotter. In any event, drastic climatic change will ensue!

As soon as I realized this unfortunate thermodynamic consequence, I hurried to dismantle my half-built accelerator. I urge others who try to construct such an earth-based accelerator to consider first the climatic consequences of a temperature rise to 12,000 degrees K. Incidentally, these considerations apply qualitatively to any new source of power that significantly exceeds what we already receive from the sun. The availability of nearly unlimited power from fusion, in particular, could confront us with the ultimate thermodynamic crisis.

RICHARD WOLFSON

Department of Physics
Middlebury College
Middlebury, Vt.

To the Editors:

Readers who were intrigued by the report on the biology of aging in "Aging Comes of Age" ["Science and the Citizen," May] will be interested to know that some of the major conclusions presented at the Santa Fe conference were anticipated more than 30 years ago. In a 1957 article in the journal *Evolution,* George C. Williams of Michigan State University presented a selectionist theory of aging and explicitly proposed that aging was the result of genetic programming caused by the action of multiple genes. These ideas, long resisted by physicians and biologists studying aging, have now achieved wide acceptance.

The centerpiece of Williams's theory is his proposal that aging results from genes that confer some benefit early in life but that cause function to decline late in life. The wild-type allele of the *age-1* locus, discovered by Thomas E. Johnson of the University of Colorado at Boulder in the worm *Caenorhabditis elegans,* is precisely this kind of gene. The gene for Huntington's disease, an adult-onset neurodegenerative disorder, may also be an example of a gene that confers an early selective advantage but that causes morbidity and mortality late in life. There is evidence that the gene increases the fertility of its carriers even though it also leads to premature death.

The probable validity of Williams's theory should give pause to venture capitalists interested in the treatment of aging. If aging is indeed the result of multiple genes and if some or all of these genes have important functions early in life, then therapeutic manipulation of the aging process may prove to be very difficult. Individuals interested in this topic would do well to read Williams's original article. It is a model of rigorous thought and clear exposition.

ROGER L. ALBIN

Department of Neurology
University of Michigan
Ann Arbor

To the Editors:

Most music historians would say that Johann Sebastian Bach had four sons who were "accomplished musicians and composers," not three, as stated in "The Blood-Brain Barrier" ["Science and the Citizen," March]. But, more important, one of these four (as well as another, short-lived son) seems to have been an emotionally troubled under achiever, and a sixth son appears to have been mentally retarded. It is also worth noting that although the Bach family had for at least four generations before the advent of J. S. supplied central Germany with hundreds of musicians, the great majority of them were run-of-the-mill musical craftsmen.

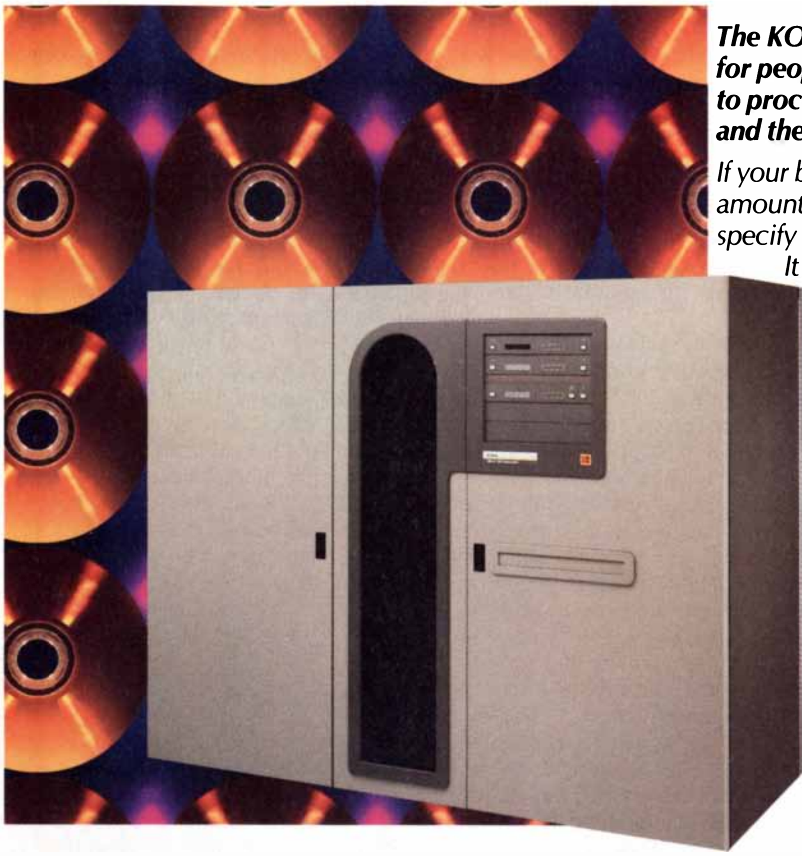
My point is that the Bach family and, presumably, other oft-cited historical families can hardly provide clear-cut evidence for either side of the nature/nurture controversy. Moreover, I wonder whether such examples are at all relevant to the relation between genetics and intelligence in the modern world. The social conditions that led to such extended families of artists (or scholars or craftsmen) in a preindustrial society no longer exist. I would imagine that all such families, regardless of their genetic peculiarities, evolved a sort of microculture not to be found in modern families and no longer accessible to direct study, which would have been implicated in the accomplishments (and failures) of its members.

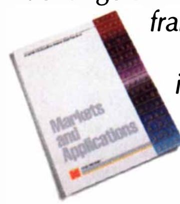
DAVID SCHULENBERG

Department of Music
Columbia University
New York, N.Y.

ERRATUM

In the top graph on page 40 of the April SCIENTIFIC AMERICAN, the right-hand scale is incorrectly labeled. The scale should read from −10 to 4 degrees Celsius, in increments of 2 degrees.




Markets
and
Applications

50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

AUGUST, 1939: "By drilling deliberately crooked oil wells instead of straight ones, a young Oklahoman has added immeasurably to the dwindling oil resources of the world. H. John Eastman, though told by engineers that it couldn't be done, showed them how drilling on curve and slant can outwit geology, put a subterranean cork in wild and flaming wells and, by angling out under the ocean, open up to man's use huge new reserves of petroleum. Where oil is found under cities, buildings, parks and cemeteries, Eastman's process can tunnel down at an angle from far enough away to make unnecessary such projects as putting derricks on the lawn of Oklahoma's state capitol."

"With wings spreading 152 feet and hulls 106 feet long, Boeing Clipper ships have inaugurated transatlantic airmail service; it is expected that, by the time this issue reaches the reader, transatlantic passenger service will be flying on regular schedule."

"British automobile traffic appears to be moving backwards. A survey of traffic speeds by Great Britain's minister of health reveals that in 1904 the average speed of a horse-drawn bus traveling from Swiss Cottage to Oxford Circus, in the heart of London, a distance of about four miles, was not quite nine miles an hour. In 1937 several tests indicated an average speed for the same journey by motorbus of about eight and a half miles an hour."

"Seedless watermelons have been produced by treating the unpollinated flowers with naphthalene acetic acid. The melons were all seedless but varied in shape. The texture of these fruits was firm and solid; no difference in flavor could be detected from normally pollinated fruits."

SCIENTIFIC AMERICAN

AUGUST, 1889: "A murderer in New York named Kemmler has been sentenced to death, and his execution will be the first under the newly enacted law by which electricity is substituted for the rope, hanging being abolished. The lawyers for the condemned man are making strenuous efforts to save the criminal on the ground that the new mode of causing death is experimental, is a failure, is simply a way to inflict grievous bodily torture, not a legitimate mode of execution."

"The phenomenon of fluid friction may be beautifully shown by a very simple experiment devised by Sir William Thomson. From a gas fixture, or other convenient support, two wires are hung, and to the lower end of each one is fastened a rubber loop. Into each loop an egg—one raw, the other hard boiled—is slipped, with its long axis vertical. The eggs are gently turned once or twice round and then let go. The boiled egg keeps twisting to and fro, after the manner of a torsion pendulum, while the raw one comes almost immediately to rest. The explanation is easy. The hard boiled egg, being rigid throughout, turns as a whole, while the raw egg, being soft inside, has only its shell moved by the torsion of the wire, the contents remaining stationary because of their greater inertia. The shell is thus made to rub to and fro on its contents, and being very light, is soon brought to rest."

"USEFUL, IF TRUE: Some one has made the discovery, or rather makes the assertion, that a fly always walks upward. Put a fly on a window, and up he goes toward the top; he can't be made to walk downward. So an inventor has made a screen divided in half. The upper part laps over the lower, with an inch space between. Well, as soon as a fly lights on the screen, he proceeds to travel upward, and thus walks straight out doors. By this means, a room can be quickly cleared of flies."

The MacCoy pneumatic tool

"The Hotel Bernina, at Samedan, Switzerland, has for some time been lighted with electricity, power being supplied by a waterfall. As during the day the power is not required for lighting and is therefore running to waste, the proprietor of the hotel has hit upon the idea of utilizing the current for cooking, and an experimental cooking apparatus has been constructed. This contains German silver resistance coils, which are brought to a red heat by the current, and it has been found possible to perform all the ordinary cooking operations in a range fitted with such coils."

"The new translucent substance intended as a substitute for glass has been adopted for some months in some public buildings of London, and various advantages are claimed for it, among these being such a degree of pliancy that it may be bent backward and forward like leather, and be subjected to very considerable tensile strain. The basis of the material is a web of very fine iron wire, this being enclosed, like a fly in amber, in a sheet of translucent varnish."

"Observation in public places gives satisfactory evidence that the use of cigarettes is rapidly on the decline. Whether this is because of the stringent laws passed in many of the States against selling them to minors, or because smokers have come to their senses and have taken warning from their own experience and the unanimous condemnation of smoking cigarettes by the medical profession, we know not."

"We illustrate in the present issue a machine which is of peculiar interest from a strictly mechanical as well as from the general point of view. The MacCoy implement involves the application of the principle of light hammer blows, repeated with high velocity, to the execution of mechanical work in wood, metals, stone, and other materials. A single cutting tool is used, but is subjected to an immense number of blows repeated with the highest velocity. It is called a pneumatic tool, as it is generally designed to be held in the workman's hand and to be driven by compressed air."

SCIENCE AND THE CITIZEN

Misprints

Seeking new standards for forensic DNA typing

On February 5, 1987, Vilma Ponce and her two-year-old daughter were found stabbed to death in their South Bronx apartment. Police arrested a neighbor, José Castro and, during the search, found a speck of dried blood on Castro's wristwatch. The blood was sent to Lifecodes Corporation in Valhalla, N.Y., one of two commercial laboratories in the U.S. that perform the new forensic technique known as DNA "fingerprinting." The laboratory reported that the DNA on the watch matched Ponce's. But in a lengthy pretrial Frye hearing—a hearing held to ascertain whether a forensic technique is accepted by the scientific community—a distinguished group of scientists, including defense witness Eric S. Lander of the Whitehead Institute for Biomedical Research and prosecution witness Richard J. Roberts of Cold Spring Harbor Laboratory, challenged Lifecodes' conclusion. Part way through the hearing, Lander, Roberts and two other scientists met and agreed, in an unusual joint statement, that Lifecodes' data were scientifically unreliable. The judge, Gerald Sheindlin, was expected to rule on the DNA evidence in July.

In the wake of Castro and half a dozen other challenges to DNA evidence now under way, many observers wonder whether the judicial system has moved too quickly to embrace DNA testing. Widely publicized as a revolutionary tool for law enforcement, DNA evidence has already led to perhaps 200 convictions (and some acquittals) in the U.S., according to estimates by Peter Neufeld and Barry Scheck, the two defense attorneys in the Castro case. Although no one disputes the power of the technique, there is concern that it is being misused and that adverse publicity could cause it to be discredited.

Observers agree that standards are urgently needed. A New York State panel is expected to call for an advisory committee to oversee quality controls and for state certification of forensic DNA laboratories. The National Academy of Sciences has proposed a study to set guidelines. The National Institute of Justice and the FBI, which had declined to fund the study, decided in June to support it after all.

DNA typing is in principle straightforward. The technique relies on the existence of DNA sequences that tend to vary in length from one person to the next. DNA is chopped up by enzymes and the fragments separated by gel electrophoresis. The fragments are then exposed to radioactively tagged probes—pieces of DNA that recognize these highly variable regions and hybridize with them. The radioactively tagged fragments show up as a pattern of dark bands on X-ray film, at positions corresponding to their sizes. From population data, investigators can calculate how frequently a particular fragment size, or allele, occurs in a population. With a sufficient number of probes, the probability that all the bands for one individual will match all the bands for another individual can be very low indeed.

In forensic DNA typing, however, the data can be difficult to interpret. At a conference held last fall at Cold Spring Harbor's Banbury Center, molecular geneticists, statisticians, law-enforcement officials and lawyers discussed criteria for deciding whether two samples match, based on visual inspection of the film and on computer systems that calculate the molecular weight of fragments by their position on the film. They agreed on the need for controls, because fragments of the same size in different lanes may migrate at different rates, resulting in what is called band shifting. Also, forensic evidence is often degraded or contaminated, which can result in smearing or in extra bands. "These bands don't look like supermarket bar codes," one scientist noted.

In the only proficiency test of forensic DNA typing yet reported, the California Association of Crime Laboratory Directors sent 51 pairs of samples to Lifecodes in 1987; the company reported 37 of the pairs correctly and called the remaining 14 inconclusive. Cellmark Diagnostics in Germantown, Md., the other commercial forensic DNA test laboratory in the U.S., was sent 49 pairs of samples; it got 44 pairs right and made one error.

Although Lifecodes performed reliably in the proficiency test, problems there came to light in the Castro hearing. To determine whether the blood on Castro's watch came from a woman, the laboratory applied a probe for a locus on the Y chromosome, but the control was female DNA; there was no male control to prove that the probe was working. The defense also challenged Lifecodes' procedure for declaring a match. In its report to the district attorney, Lifecodes stated: "Fragments with measurements that are within approximately 2 percent [in molecular weight] of each other ([a] 99.7 percent confidence level) are considered indistinguishable." But at two sites, the sizes calculated by Lifecodes' computer differed by more than 2 percent in DNA from the watch and from Ponce. Michael L. Baird, director of Lifecodes' paternity and forensics laboratory, testified that the company had in this case relied on subjective visual comparison rather than on its own objective standard to declare the match.

Baird, who would not comment on the Castro case, maintained in a telephone interview that Lifecodes screens data visually and then turns to the computer sizings to declare a match. The numbers are first corrected for band shifts, which are determined by the position of several control bands; numbers that fall within a range of 2 percent are considered matches, he said. Such control bands, however, were not used in the Castro case. Baird also said that when there is enough forensic evidence, part of it is mixed with DNA from the suspect and run on the gel alongside the individual samples. This procedure is widely advocated as one of the best methods to control for differences in band size: if the DNA's are identical, the mixed sample will produce the same pattern as the unmixed ones.

Scientists familiar with several Lifecodes cases say they are not convinced that the criteria for declaring matches are always so objective. In one death-penalty case now under way

in Marietta, Ga., Kevin C. McElfresh, a scientist at Lifecodes, acknowledged that the films from the rape-murder suspect and the forensic evidence did not match visually but maintained that "there is... a consistent nonalignment of bands throughout the test, telling us there's a match." Lander says Lifecodes presented no internal controls to prove a shift had occurred.

The Castro hearing also saw a dispute over whether the population in question—Hispanic, in the Castro case—had a distribution of alleles that would result from random mating, a state known as Hardy-Weinberg equilibrium. If the alleles have not been randomly assorted, then the probabilities of matches between individual bands cannot simply be multiplied to give the overall probability of a match. Another issue of population genetics arose in a Texas rape-murder case that took place in a small, heavily inbred black community. The analysis by Lifecodes ignored this fact, Lander said.

Some scientists say that many of the procedures for which Lifecodes was criticized in the Castro hearing have since been modified. Simon Ford, a molecular geneticist at the University of California at Irvine, noted that "the Frye rule states that a technique must be shown to have gone from the experimental to the demonstrable stage. Cellmark and Lifecodes are both modifying their procedures based on these challenges. What better proof that they are still experimental?" Roberts, who is skeptical of the courts' ability to sort out scientific truths, says if national guidelines could be set "the onus would then be on the companies to explain failures to meet them." —*June Kinoshita*

Big Brother, Ph.D.

Should scientists or bureaucrats form science's police force?

Can the traditional mechanisms for keeping science honest—verification by repetition of observation or experiment, peer review and trust in a worker's reputation—maintain the health of an enterprise founded on intellectual honesty? Or do the changing conditions for research—large sums of money, team research and media celebrity—mean new mechanisms have to be fashioned? Events have already begun to answer the question in ways that make some workers uneasy.

Last spring, a congressional subcommittee spent two days cross-examining biologists and hearing from secret-service agents called in to analyze laboratory notebooks. The hearings, conducted with all congressional pomp, were rooted in charges by a postdoctoral researcher that results had been reported falsely in an immunology paper published by David Baltimore of the Whitehead Institute for Biomedical Research and others three years ago in *Cell.* The charges led to a string of investigations; the most recent one found "serious" errors in the paper, but no evidence of deliberate wrongdoing. Baltimore maintains that the errors were innocent and says there is no reason to doubt the paper's main conclusions, but yet another inquiry is now under way.

The ostensible subject of congressional interest is how science deals with whistleblowers: the accuser says she could not get a research appointment because of her charges. John D. Dingell, Democrat from Michigan, who is leading the inquiry, maintains Congress has to ensure fair treatment for those who charge that federal research funds are being ill spent.

Scientists fear that political pressures for greater regulation of science could stifle creativity. "We are primarily concerned that excessive publicity will lead to unwise regulation or the establishment of an overintrusive or expensive bureaucracy," says Frank Press, president of the National Academy of Sciences: "The kinds of things that would worry me are establishing intrusive methods of policing laboratories, such as subpoenaing of laboratory notebooks."

Press is concerned that nonscientists do not fully appreciate the distinction between honest error, which is a part of science, and deliberate fraud. "In terms of the number of cases and their impact on science, fraud is not a big issue," Press believes. Nonetheless, he agrees that the notoriety of cases of scientific fraud—usually understood as deliberate misconduct—means science cannot ignore them. Recently, paleontologists were shocked by charges made by John A. Talent of Macquarie University in Australia against Vishwa Jit Gupta of Punjabi University. Talent alleges that in 300 scientific reports published over a 20-year period, many with prominent and unsuspecting coauthors, Gupta has provided false accounts of where he found fossil assemblages, thus corrupting a whole body of literature.

Press argues that such gross departures from acceptable practice would normally be found out by science's own in-built mechanism: the inability of other researchers to validate a false scientific claim. Nonetheless, a recent study published in the *Journal of the American Medical Association* found that 7 percent of clinical trials of new drugs supported by the Food and Drug Administration had shortcomings "ranging from negligence to fraud" that undermined their conclusions. Biomedical research seems particularly prone to misconduct.

Some changes seem inevitable. Ephraim Racker, a biochemistry professor at Cornell University, has suggested a preemptive strategy: researchers should present Congress with a plan for what science can do to police itself, a plan that should define a role for universities and department heads. Drummond Rennie, deputy editor of the *Journal of the American Medical Association,* believes scientists should conduct data audits themselves in order to forestall hasty congressional action.

A recent study by the Institute of Medicine, which is affiliated with the National Academy of Sciences, gave its blessing to an office of scientific integrity recently established by the National Institutes of Health in response to the new concerns. John Maddox, editor of the British journal *Nature,* is skeptical: "What's needed is not to make the NIH more of a policeman but to make universities and institutions more responsible for what their people do," he says. Maddox has frequently argued the importance of peer review of scientific results prior to publication—the traditional safeguard. The public announcement of cold fusion by B. Stanley Pons and Martin Fleischmann (see "Morning After," page 12D) is seen by many researchers as a case in point. A thorough peer review of Pons and Fleischmann's research report prior to its publication, instead of the rushed treatment it received, might have flagged faulty measurements and avoided enormous wasted scientific effort.

Peer review can help to catch honest errors but is unlikely to entrap workers who commit deliberate fraud. Research institutions are increasingly establishing procedures for dealing with allegations of such serious misconduct as well as lesser lapses. According to Brian W. Kimes, acting director of the new office of scientific integrity at the NIH, the office will monitor institutional investigations but intervene only if they break down. "I'd hate to think we were going to create rules that would destroy the creativeness of individuals," Kimes says. The office

will also seek to educate institutions about how to encourage good scientific practice, such as keeping original data accessible. "The NIH will be taking a stronger and stronger position with regard to data access and data storage," Kimes says. Scientists are bracing themselves. *—Tim Beardsley*

PHYSICAL SCIENCES

Surfing Photons

They ride a wave of electrons to attain X-ray energies

Imagine an ocean liner forging a path across a peaceful bay. On the shore nearby a surfer searching for the ultimate wave gets an awesome vibe. Paddling out to the ship's wake, the surfer catches the artificial wave and rides it across the bay.

Surfing on the wake illuminates the novel physics that underlies a proposal to convert short pulses of laser light into intense, focused bursts of X-radiation. The emerging technology, as conceived by Scott C. Wilks and John M. Dawson of the University of California at Los Angeles, could develop into a more efficient, powerful alternative to free-electron and X-ray lasers. The new technique involves directing a particle beam through a sea of electrons, thus generating a powerful wake that would propel surfer-photons to X-ray energies.

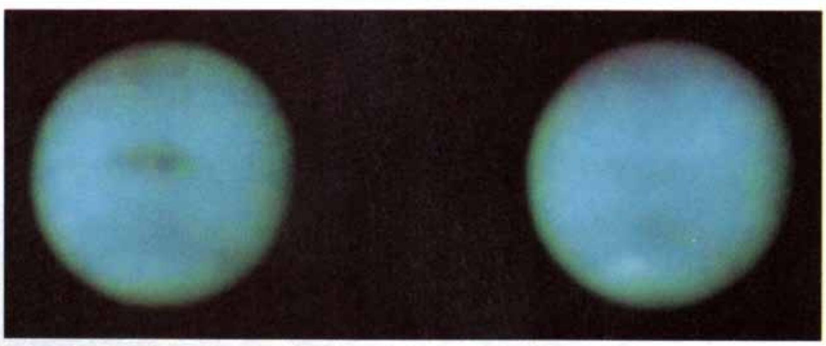
To create the sea of electrons, a gas—perhaps enclosed within a long chamber—is heated to a temperature at which electrons are stripped from atomic nuclei: a state known as plasma. A beam of particles traveling at nearly the speed of light is then directed into the plasma; a wave of electrons follows in its wake. This electron wave can be thought of as a wall of negative charge moving in the same direction and speed as the beam.

According to Wilks and Dawson, a pulse of laser light at a particular frequency can be released into the plasma behind the particle beam and just in front of the rolling wall of negative charge. The laser pulse moves through the plasma as a packet of photons. The wall of negative charge propels the photons; that is, the photons steadily absorb energy from the plasma wave. Thus, the workers say, the photons are accelerated, or gain energy, and leave the plasma as a burst of high-energy radiation.

Wilks, Dawson and their colleagues report in *Physical Review Letters* the results of their computer simulation of this photon accelerator. Given the current state of laser and plasma technologies, the workers predict that such an accelerator could increase the energy of a photon by a factor of 10, transforming violet light into X rays. The group claims that lengthening the time during which the photons interact with the plasma may amplify the photons by an even greater factor. In theory, the accelerator should transfer energy to photons more efficiently than current free-electron and X-ray lasers. The photon accelerator could generate radiation that might be used as a weapon, a surgical instrument or a precipitator of nuclear fusion.

For the first tests of the photon accelerator, the workers plan to adapt a device that accelerates electrons in plasma (see "Plasma Particle Accelerators," by John M. Dawson; SCIENTIFIC AMERICAN, March). The surf is up along the California coast, and the next few months will tell if the physicists can catch the wave. *—Russell Ruthen*

Voyager 2 *nears its final solar-system rendezvous before heading for interstellar space*



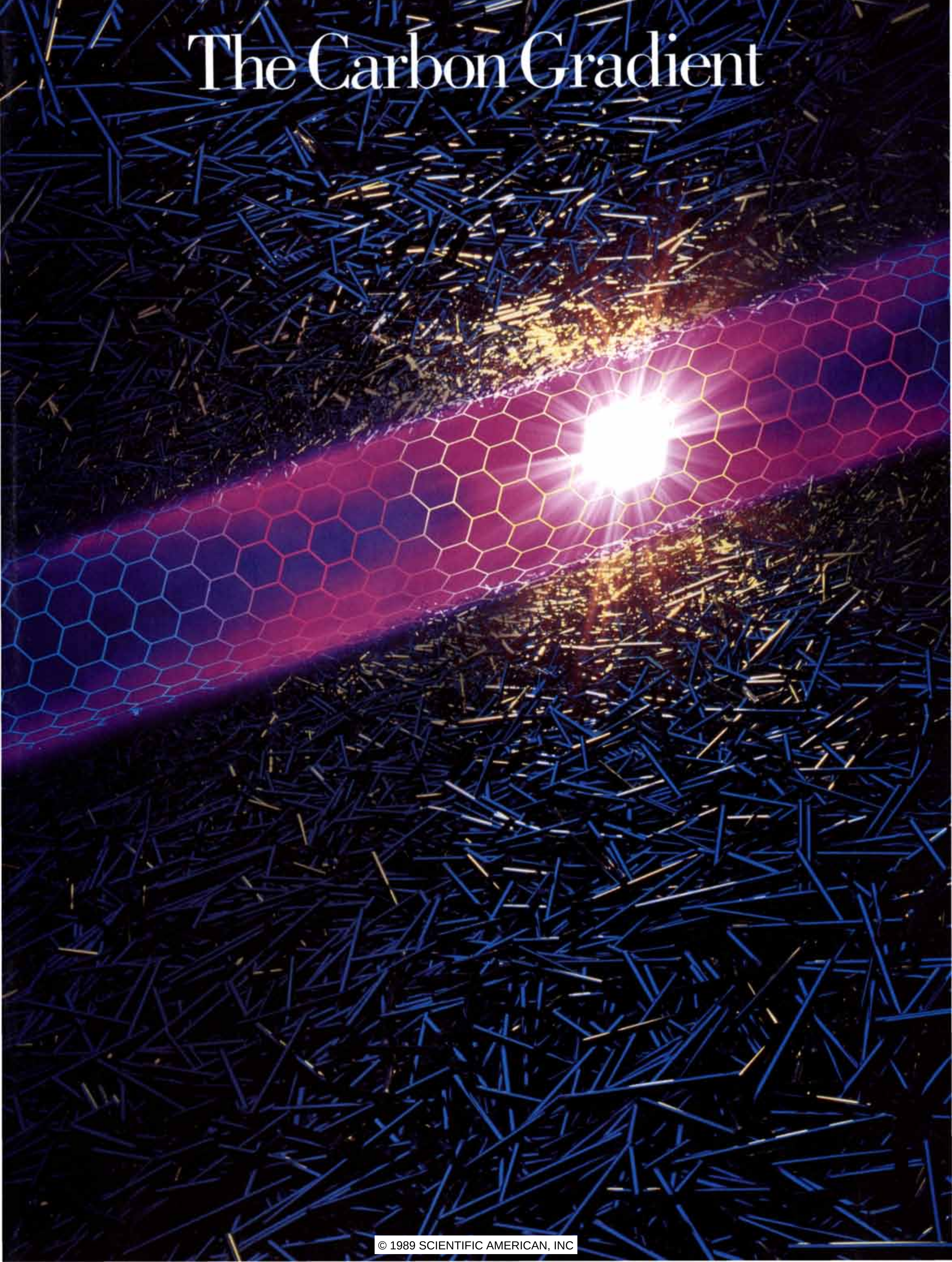
Blue-green Neptune looms out of the blackness of space into the digitized optical ken of the *Voyager 2* spacecraft, in late April. The spacecraft was then still 109 million miles out from the planet. The images confirm expectations, according to *Voyager*'s project scientist, Edward C. Stone, that the planet has an active atmosphere driven by an internal heat source. The large dark spot visible in the left-hand image, first seen in January, may be a semipermanent feature of Neptune's atmosphere, Stone speculates—perhaps an anticyclonic storm akin to Jupiter's famous red spot. By the time the right-hand image was recorded five hours later, the planet had rotated 100 degrees, bringing into view the white spot visible near the south pole. This new feature could be a thunderhead, Stone suggests; a similar but brighter white spot eventually faded. The general blue-green cast can be seen from the earth and is caused by methane. *Voyager*'s closest approach to Neptune in August should produce far more detailed images. *—T.M.B.*

Morning After

Doubts on one cold-fusion claim intensify; another gains ground

Three frenzied months later, the cold-fusion-power party seems to be winding down. Literally hundreds of investigators all over the globe have failed to validate the astounding claim announced in March at the University of Utah by Martin Fleischmann and B. Stanley Pons: that they had produced heat from nuclear fusion at room temperature, a result that could point the way to an inexhaustible source of clean energy. The claim by Fleischmann and Pons was based on running an electric current through heavy water using a palladium cathode: the experimenters said their apparatus liberates heat from the fusion of heavy hydrogen atoms that the current drives into the palladium. Yet, even though a few workers claim to have found unexplained production of heat, the effect is peculiarly elusive and there is scant evidence that fusion is responsible.

If fusion were producing heat, it would also generate by-products such as helium or tritium (an extra-heavy hydrogen isotope). Copious emission of neutrons would also be expected, which would produce gamma rays. No experiments have shown such evidence convincingly—although Nathan


The Carbon Gradient

The Carbon Gradient

Hollow carbon filaments catalytically produced by submicron-size iron particles can be the template for larger carbon fibers used in composite structural materials. A scientist at the General Motors Research Laboratories has identified how these filaments grow and why they take their characteristic form.

Carbon Fiber Growth

Fig. 1

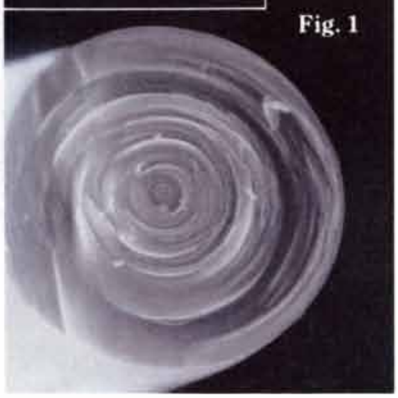


FIGURE 1: Scanning electron micrograph of a cross section of a vapor-grown carbon fiber.

FIGURE 2: Typical carbon filament grown from natural gas by an iron catalyst particle.

FIGURE 3: Schematic model showing inner and outer radii, the precipitation interface, and the nested basal planes of the outer surface.

Dr. Gary Tibbetts was measuring the diffusion rate of carbon in iron when his carefully planned experiment took an unexpected turn. Dr. Tibbetts, a physicist at the General Motors Research Laboratories, had been introducing carbon to the inside surface of a hot stainless steel tube while extracting carbon from the outer surface.

At the end of one particular trial, he found the inside surface covered with a mass of black "whiskers." His initial investigations verified that the fibers were made of carbon and that they had characteristics typical of the crystal structure of graphite. But the question of how they formed was not so easily answered. The search for an answer would change the course of his investigation and dominate his research for the next ten years.

Fig. 2

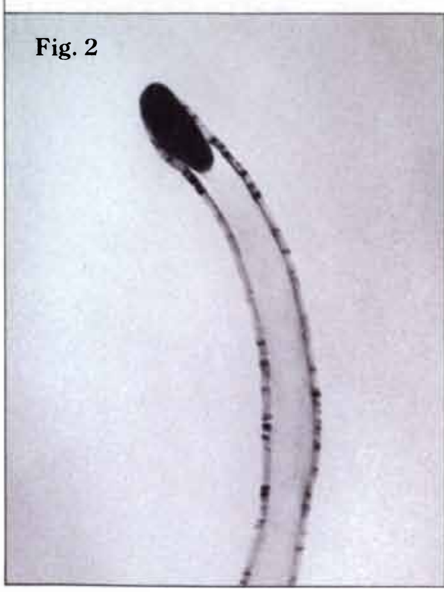
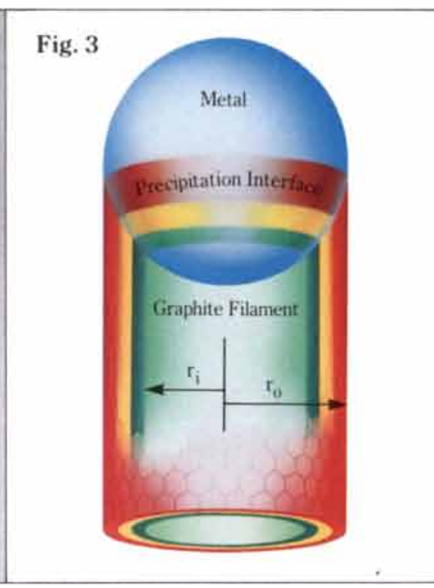


Fig. 3



The fibers that surprised Dr. Tibbetts were made up of concentric layers primarily composed of basal (0001) plane graphite, resembling in cross section the annular rings of a tree (Figure 1). Research showed that they were formed by vapor deposition of carbon on a hollow central filament. The central filament itself was grown by catalytic action on a small metal particle (Figure 2).

These long, slender, uniform filaments had been widely observed since the availability of the electron microscope. Still, no valid explanation had been advanced to account for their hollow structure. Many scientists thought that surface diffusion of carbon-containing molecules around the catalytic particle caused the hollow core.

Instead, Gary Tibbetts proposed a model in which carbon atoms from decomposing hydrocarbons diffuse through the bulk of the catalytic particle and precipitate as graphite in the growing filament. The diffusion process is driven by the carbon gradient–the difference between carbon concentrations at the adsorbing surface of the particle and at its opposite, precipitating surface (Figure 3).

The exterior surfaces of these carbon cylinders expose the basal plane of graphite because the (0001) plane has a surface free energy at 970°C of about 77 erg cm^{-2}, while a typical surface perpendicular to the basal plane has a surface energy in excess of 4000 erg cm^{-2}. The free energy required for filament growth, therefore, will be a minimum when the exterior surface is made up of basal planes–as observed in these filaments.

The entire filament, then, should consist of nested, rolled-up basal planes of graphite. Bending these planes into cylinders, however, requires that extra elastic energy be provided during the precipitation process. The core is left hollow because too much energy would be required to bend the planes near the axis into very small diameter tubes.

In describing the total energy necessary for filament formation, Dr. Tibbett's model takes into account the chemical potential change ($\Delta\mu_o$)when a carbon atom precipitates from the dissolved phase, as well as the energy required to form the surface, plus the energy needed to bend the basal planes into nested cylinders.

The change in chemical potential ($\Delta\mu$) driving the precipitation can be expressed as follows:

$$\Delta\mu = \Delta\mu_o - \frac{2\sigma\Omega}{r_o - r_i} - \frac{E a^2 \Omega}{12(r_o^2 - r_i^2)} \ln(r_o/r_i)$$

where σ is the energy required to form a unit area of (0001) graphite; Ω is the volume of a carbon atom in graphite; r_o and r_i are the outside and inside radii of the filament, respectively; E is the filament modulus; and a is the interplanar spacing.

A filament catalyzed by a particle of radius r_o will adjust its r_i to give the largest $\Delta\mu$–in fact, r_i may be directly calculated by maximizing $\Delta\mu$. Doing so yields results that compare nicely with experimental values.

Understanding the growth of the hollow core of the filaments was one key to producing them in abundance. "From there," says Gary Tibbetts, "it is a simple step to thicken the filament into a macroscopic fiber by vapor deposition of carbon on the exterior surface. The deposited carbon has a high degree of orientation parallel to the tube axis, giving the fiber exceptional stiffness.

"Fibers of this type should be excellent for making chopped-fiber composites using plastic, ceramic, metal, or cement matrices. GM's Delco Products Division is already building a pilot plant to develop a low-cost production process that would permit the use of vapor-grown fibers in high-volume applications."

General Motors

MARK OF EXCELLENCE

THE MAN BEHIND THE WORK

Dr. Gary G. Tibbetts is a Senior Staff Research Scientist in the Physics Department of the General Motors Research Laboratories.

Gary received his undergraduate degree in physics from the California Institute of Technology. He holds both an M. S. and a Ph. D. in the same discipline from the University of Illinois.

Dr. Tibbetts joined General Motors after two years of postdoctoral work as Guest Scientist at the Technical University of Munich. Since coming to the Labs in 1969, Gary has pursued interests ranging from carbon filaments, to surface physics, to chemical vapor deposition. He has published almost forty papers on the results of his research.

Gary is a member of the American Physical Society, the American Carbon Society, and the Materials Research Society. In 1988, he was a GM Campbell Award Winner. Gary lives in Birmingham, Michigan, with his wife and their three daughters.

S. Lewis of the California Institute of Technology, a prominent skeptic, reports finding chemical effects that might deceive the unwary. "No group has verified all the essential features of the Utah claim," Lewis observes.

Among the few investigators who continue to see evidence of anomalous heat production are Robert A. Huggins and his colleagues at Stanford University and several independent groups at Texas A&M University. Huggins says his studies suggest that most investigators are "wasting their time" because they use too small a current. He has not yet searched for neutrons or other fusion products.

"We have still not eliminated all possible trivial chemical explanations, but we are rapidly approaching that point," says A. John Appleby of Texas A&M University, whose microcalorimetry experiments also seem to detect excess heat. Nevertheless, tests for helium have turned up nothing. Several laboratories have confirmed the presence of unusual amounts of tritium in heavy water from experiments run by Kevin L. Wolf, also of Texas A&M, who reports detecting neutrons being emitted at a low rate. Wolf believes the palladium has to be recast many times in order to drive out all traces of ordinary hydrogen, which he maintains "poisons" the system. "We are completely confident it is tritium," Wolf declares. "The only loose end is—how did it get there?"

Most workers suspect the answer is contamination. Even Fleischmann and Pons no longer claim to detect neutrons and helium-4 production in their experiments—observations they had offered as evidence for fusion. Richard D. Petrasso and others at the Massachusetts Institute of Technology have damningly criticized Fleischmann and Pons's gamma-ray measurements, which were presented to show the emission of neutrons with precisely the energy expected from fusion. Petrasso's critique suggests that the gamma-ray "signal" was a glitch that actually appeared at a clearly different energy. The criticism poured on Fleischmann and Pons intensified when Los Alamos National Laboratory announced in June that a planned collaboration with the Utah researchers had fizzled because the University of Utah was dragging its feet.

In spite of the mounting skepticism over cold fusion as a source of energy, a quite separate claim by Steven E. Jones of Brigham Young University that he has detected fusion at an extremely low level in a somewhat similar experiment has received some support. "The excess heat claimed by Fleischmann and Pons is almost certainly not due to fusion," Jones says. Still, he maintains that "things have strengthened dramatically for cold fusion" at the level of a few fusions per minute—which produces no detectable heat but a small flux of neutrons. Jones's claim is apparently supported by work done by Howard O. Menlove and his colleagues at Los Alamos and by work at the University of Bologna. Menlove and Jones believe they have observed statistically significant production of neutrons in replication experiments conducted at Los Alamos.

Menlove also reports detecting single neutrons and "bursts" of hundreds of neutrons being emitted in an experiment first done by F. Scaramuzzi at the Center for Energy Research in Frascati, Italy. Heavy hydrogen is forced under pressure into titanium shavings, and the temperature is cycled.

Even low-level fusion is not universally accepted. Notably, Moshe Gai of Yale University reports that his replication of Jones's experiment failed to produce very low neutron levels. Norman Hackerman, president emeritus of Rice University, and J. Robert Schrieffer of the University of California at Santa Barbara, who were cochairmen of a government-sponsored conference on cold fusion held in Santa Fe, consider the case for Jones's low level of fusion to be, at best, "moderate." Schrieffer, a Nobel laureate and senior statesman of physics, agrees with Hackerman that the evidence for anomalous heat production is "much more tenuous." Hackerman notes that those claiming heat production cannot always produce the effect. In the meantime, the Energy Research Advisory Board has established a special panel on cold fusion, which has started laboratory visits. Many months' work may be needed to dispel fusion confusion. —*T.M.B.*

Pride and Prejudice

Cosmic lithium suggests the universe is open

Current cosmological prejudice requires that the universe be flat—just balanced between a universe doomed to recollapse and one destined to expand forever. The cosmos, however, does not appear to be swayed by prejudice. A recent paper in *Physical Review Letters* by a team at Yale University provides the latest evidence suggesting either that the universe will expand forever or that the bulk of its mass is hidden in exotic, undetected particles.

For several decades cosmologists have been almost certain that light isotopes such as deuterium, helium and lithium-7 (^{7}Li) were synthesized in the first few minutes after the big bang. The amount of these isotopes produced in cosmic nucleosynthesis depends critically on the density of baryons—or neutrons and protons—in the universe at that time. All theoretical nucleosynthesis studies show that to produce the observed amounts of helium and deuterium the baryon density must be much less than the critical value, the density needed to transform the universe from open (ever expanding) to closed (doomed to recollapse).

^{7}Li can in principle also be used to determine the baryon density, but its abundance is miniscule compared with the other isotopes. On the experimental side this has made accurate observations of ^{7}Li difficult; on the theoretical side it means that the ^{7}Li abundance is particularly sensitive to uncertainties in the known nuclear-reaction rates.

The Yale team's approach to the problem is somewhat unusual; it begins not with nucleosynthesis but with stellar evolution. In 1982 François and Monique Spite observed that the amount of lithium in the oldest stars of the galaxy appeared to approach a constant value. The Spites conjectured that this value was the primordial stellar abundance, left over from cosmic nucleosynthesis.

The Yale team's first step was to use some of the most sophisticated stellar-evolution models in existence to retrodict the primordial lithium abundance. They confirmed the Spites' prediction that the lithium in old stars asymptotically approaches a constant—the presumed primordial abundance.

The next step was to check what baryon density produces this value in a nucleosynthesis calculation. Here, too, the procedure was slightly unusual. Since the lithium abundance is sensitive to the various parameters that go into the reaction rates, the Yale team varied the parameters randomly, taking a Monte Carlo approach. The result, says Lawrence M. Krauss, one of the investigators, "is probably the first nucleosynthesis calculation with statistical confidence limits."

What did the group find? The results depend on the assumed value of the Hubble constant, which measures the expansion rate of the universe. If one chooses 50 as a value that many astronomers would accept, the two-standard-deviation lower limit to the baryon density of the universe is about 2 percent of the critical density, and the upper limit is about 10 percent.

These numbers are not new but support previous claims that the universe cannot be closed by baryons alone. Theorists who claim that the universe is exactly flat are, once again, forced to invoke nonbaryonic matter to do the trick. —*Tony Rothman*

Coming Down in Sheets

A model mantle re-creates motifs of plate tectonics

When one of the earth's surface plates reaches the end of its geologic lifetime, it warps downward along a linear subduction zone and descends sheetlike into the mantle. Yet when hot material rises from the mantle—the 3,000-kilometer-thick shell of solid but ductile rock that lies between the crust and the metallic core—it often comes up as a focused plume. Why should upwelling and downwelling in the mantle look so different?

A computer model of the mantle now suggests that downgoing sheets and rising plumes are natural consequences of the mantle's spherical shape. The model, developed by Dave Bercovici and Gerald Schubert of the University of California at Los Angeles and Gary A. Glatzmaier of the Los Alamos National Laboratory, attempts to reproduce the convective circulation of the mantle. Heat from the core or from the decay of radioactive elements in the mantle drives the convection: hot rock rises buoyantly and circulates horizontally, driving the movement of surface plates, until it cools and sinks at a subduction zone.

It had been assumed that the shape of subduction zones reflects the rigidity of the surface plates, which would cause them to bend along a line and retain their sheetlike form as they descend. The model's results, reported in *Science*, now suggest that subduction might look the same even in the absence of rigid plates.

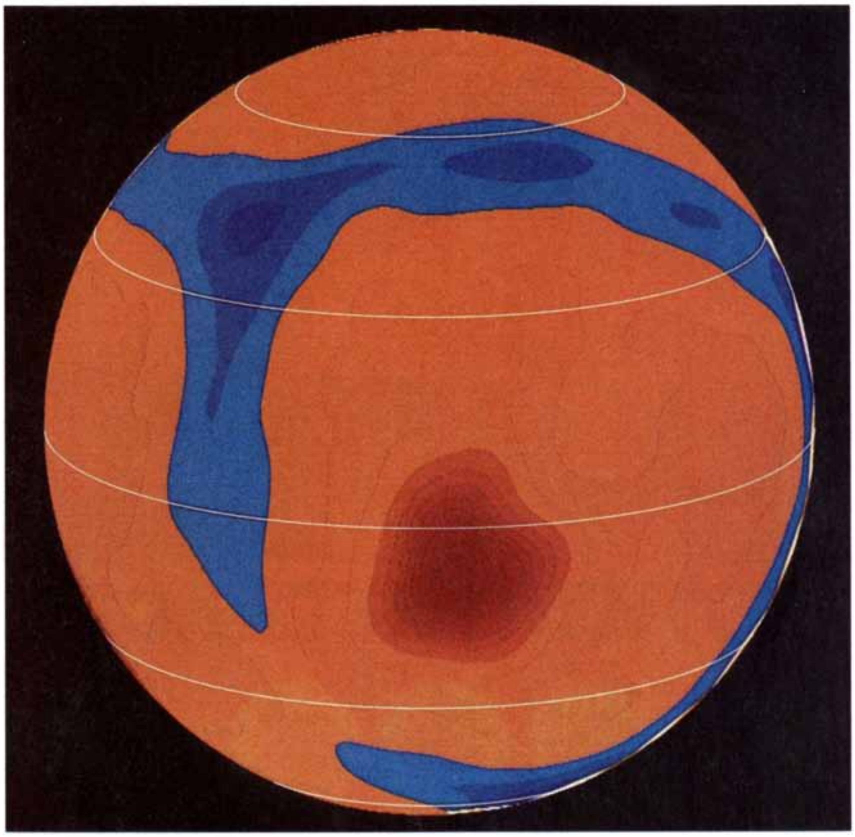
Earlier attempts to model mantle convection could not say much about the flow's geometry because, for the sake of simplicity, they generally reduced the mantle to a flat-bottomed box or a single vertical plane. In general such models also assumed a uniform rock density instead of a density that increases with depth. Bercovici and his colleagues, exploiting the power of supercomputers and building on a model Glatzmaier had developed earlier for studying convection in the outer layers of the sun, constructed a mathematical model that simulated convection in a three-dimensional spherical shell with a density stratification similar to that of the real mantle. (A model devised earlier by John Baumgardner of Los Alamos had also simulated mantle convection in a spherical shell.)

Even so, the model was highly simplified: it lacked a surface layer of rigid plates, it assumed less active convection than takes place in the real mantle, and it took rock viscosity to be constant, even though the viscosity actually varies both radially and laterally. In addition, since the ratio of the heat that flows into the mantle from the core to the heat generated within the mantle is not known, the workers ran several model versions, assuming different ratios of core to internal heating. They observed a persistent pattern of cold sheets plunging to the very base of the convecting layer and hot plumes welling up.

Seismic data confirm that sheetlike structure is preserved to great depths in subduction zones, a finding the authors say is hard to explain on the basis of the plates' rigidity alone. Geophysical data also support the suggestion that mantle upwelling takes place exclusively in plumes. Even though mantle material does rise in sheets at the midocean ridges—linear rifts between diverging plates—the hot, seismically slow material seems to originate near the surface, welling up as a passive response to the plates' divergence. Deep-seated plumes such as the ones underlying such "hot spots" as Hawaii and Iceland seem to be the normal form of upwelling.

Just why should the pattern of descending sheets and rising plumes arise naturally in a convecting spherical shell? Bercovici and his colleagues are analyzing their results

A computer simulation reproduces the style of heat-driven circulation in the earth's mantle



***RADIAL FLOW** in the mantle was traced by a computer model that treated the mantle as a spherical shell of convecting fluid. Focused upwellings* **(red)** *and sheetlike downwellings* **(blue)** *appeared consistently in the evolving flow pattern. Gary A. Glatzmaier provided the image.*

in search of an answer. The workers also hope to build more realistic models that would take into account the real mantle's varying viscosity; such models might reproduce additional features of plate tectonics—perhaps even the choreography of the plates themselves. —*Tim Appenzeller*

BIOLOGICAL SCIENCES

Successful Sex

A new theory could explain why women must put up with men

Why do most higher animals and plants bother with sex? On the face of it, asexual reproduction has a big advantage: every member of an asexual population can produce offspring, whereas only half of the members of a sexually reproducing population—the females—can do so. Asexual populations can therefore multiply twice as fast as sexual ones. Indeed, several species of lizards now exist only as females that reproduce parthenogenetically. Yet sexual reproduction persists in spite of its twofold disadvantage, and biologists have been unable to agree on a general explanation for its success.

A new theory has been presented in *Nature* by Mark A. Kirkpatrick and Cheryl D. Jenkins of the University of Texas at Austin. They propose that sexual populations outcompete nonsexual ones because sex enables a common type of favorable mutation to become established more quickly. This type of advantageous mutation is most beneficial when an organism carries it in a double dose: that is, when both of an individual's copies of a gene carry the mutation. If only one copy carries it, the benefit is reduced.

Some genetics. In a sexual population, each parent contributes just one copy of each gene to each offspring. It follows that once a favorable mutation arises, random mating will soon produce individuals fortunate enough to inherit two copies of the mutation, thereby gaining the maximum benefit.

In an asexual population, however, parents pass on both copies of each of their genes to offspring. As a result asexually produced offspring fare less well: they cannot inherit two copies of the mutation and so reap the full reward until the mutation happens to occur a second time. Meanwhile, the sexual competitors are enjoying the maximum benefit. Kirkpatrick and Jenkins show that this benefit outweighs the twofold reproductive disadvantage of sex under many lifelike conditions.

"I personally think it's a pretty good idea that as far as I know everyone has missed," comments James J. Bull, also at Austin. Many explanations for sex have relied on the notion that the shuffling of different genes that occurs during sexual reproduction might confer advantage by bringing together favorable clusters of genes. Kirkpatrick and Jenkins, however, require only single pairs of genes for their explanation. —*T.M.B.*

Bloodless Coup

Inhibiting blood vessel growth could starve solid tumors

Carl W. White and his colleagues at the University of Colorado Health Sciences Center in Denver recently treated a young boy afflicted with a rare but dangerous condition: excessive growth of capillaries in the lungs. The disease responded to interferon, which in addition to being a regulator of the immune system apparently inhibits blood vessel growth. The treatment may have been the first based on inhibiting angiogenesis (blood vessel growth), a strategy that could have broad applications.

A malignant tumor, for example, needs an abundant blood supply in order to grow; proliferating cells require more oxygen and produce more waste than ordinary tissue does. Indeed, many cancers reach detectable size only after they have developed a blood supply, notes Judah Folkman of the Harvard Medical School, well known for his studies of angiogenesis. Recent experiments by Folkman, described in *Nature*, suggest that acquiring the ability to stimulate angiogenesis is a critical early step in the development of malignancy. Other studies, meanwhile, are expanding knowledge about substances that inhibit the process.

Folkman, together with Douglas A. Hanahan of the Cold Spring Harbor Laboratory and other investigators, has studied cancer development by exploiting a recently devised tool: genetically engineered animals prone to a particular type of cancer. Specifically, the workers studied tumor development in a strain of mouse that expresses a cancer-causing gene in the insulin-producing cells of the pancreas. Those cells therefore frequently become cancerous.

In addition to tumors, the pancreases of the engineered mice contain precancerous islets of "hyperplastic" cells—cells that grow somewhat faster than normal but that have not fully escaped from growth controls. The workers were able to extract and culture these hyperplastic islets as well as the tumors and grow them along with endothelial cells in a gel in the laboratory. The investigators found that tumors reliably induced nearby endothelial cells to form would-be capillaries, thus mimicking a blood supply. (Presumably, the tumors were releasing substances that promote blood vessel formation.) Even more interesting, the investigators found that a small proportion of the hyperplastic islets could do the same thing.

The number of hyperplastic islets able to induce growth of would-be capillaries increased sharply after seven weeks in culture. Most important, the workers found that the number of tumors able to do the same thing increased in parallel. This finding strongly suggests that stimulating a blood supply—perhaps as a result of a genetic change—not only is important for tumor growth but actually is one of the key early events that marks transformation of a hyperplastic islet into a malignant tumor.

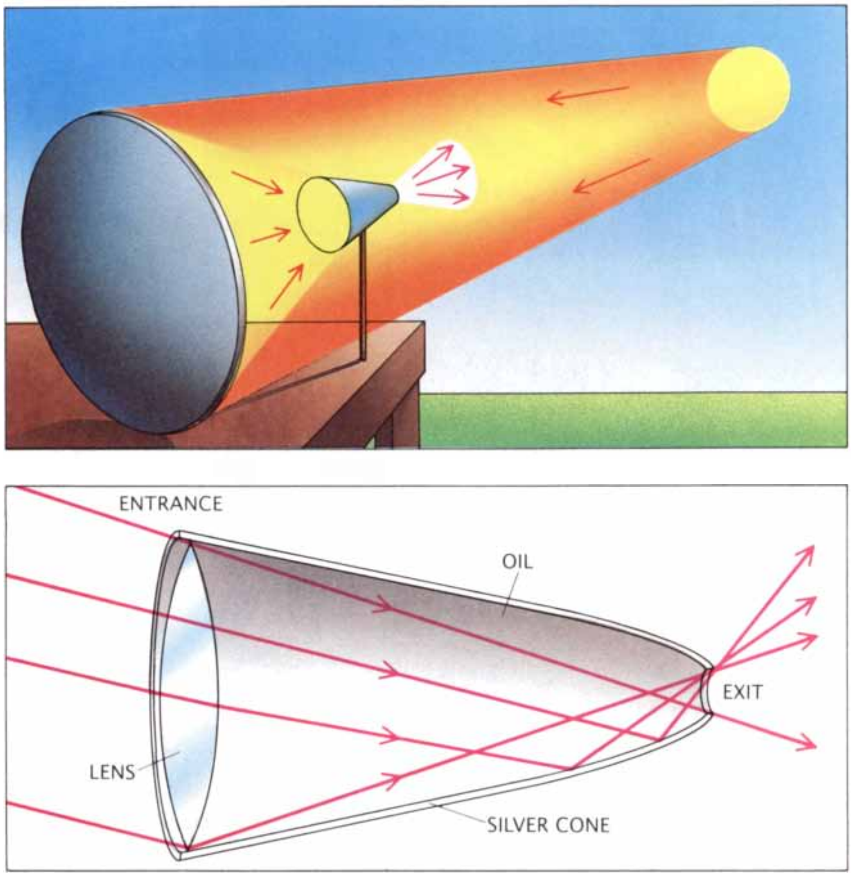
If Folkman is right, inhibiting blood vessel formation could be one more way to attack a developing tumor as well as other conditions that involve angiogenesis, among them neovascular glaucoma and rheumatoid arthritis. Folkman has recently described in *Science* how some complex synthetic carbohydrates can potentiate 100- to 1,000-fold the activity of steroids that inhibit blood vessel formation. He notes that eyedrops containing such inhibitors have already been shown to prevent blood vessel growth in the cornea, which sometimes causes blindness. —*T.M.B.*

Superantigens

Toxins may subvert the immune system's battle plans

Toxins secreted by bacteria may be responsible for many of the unpleasant and even life-threatening symptoms of infections, such as the rash, fever and water imbalance between tissues in toxic-shock syndrome and possibly the nausea and diarrhea associated with food poisoning. Recent research suggests that some potent bacterial toxins achieve their dangerous effects by overstimulating the immune system, thereby

Sunshine can be concentrated to a brightness greater than that of the sun's surface



NONIMAGING DEVICE, built by Roland Winston and his colleagues at the University of Chicago, is shown in schematic form. At the top, light from a parabolic mirror is directed into a hollow cone. The cone **(below)** ***concentrates light rays at its narrow end by the "edge ray" principle.***

unleashing immunological mayhem.

Most foreign proteins activate a smoothly orchestrated immune response. Initiation of the response involves a family of immune system cells called *T* cells. When *T* cells encounter a foreign protein (which must be presented to them by other patrolling immune system cells), the few that can bind the protein begin to divide. In this way the number of like *T* cells that can bind the protein gradually increases. Some of the *T* cells kill infected cells; others produce substances that stimulate other immune system cells, including cells that generate antibodies.

Since the variety of possible intruder proteins is enormous, an enormous number of different *T* cells are needed to ensure that any possible protein can be bound. The *T* cells achieve their variety by displaying surface receptors composed of an extensive alphabet of molecular segments arranged in different combinations. Normally a foreign protein—an antigen—will react with only a few different *T*-cell receptors, so that initially only a very small percentage of *T* cells are capable of engaging a particular interloper.

Now John W. Kappler, Philippa C. Marrack and their colleagues at the Howard Hughes Medical Institute and the National Jewish Center for Immunology and Respiratory Medicine in Denver have identified a new class of antigens that have a far broader effect on *T* cells. Produced by staphylococci (and probably other bacteria), these "superantigens" instantly fire up large classes of *T* cells, which causes them to start dividing.

The investigators write in *Science* that they exposed several classes of human *T* cells varying in a receptor subunit known as V_β to nine different staphylococcal toxins. They found that each toxin affected broad classes of *T* cells in different ways, rather than activating just a few *T* cells as do normal antigens.

Since staphylococci cause toxic-shock syndrome as well as food poisoning, Marrack suggests that the symptoms of these and other diseases, including some autoimmune diseases, might be the work of superantigens. A possible explanation is that the armies of suddenly mobilized *T* cells produce large quantities of substances that in small amounts stimulate the immune system but that in excessive quantity instead cause fever and toxic-shock symptoms. Kappler and his colleagues speculate that superantigens might help explain why some people are more susceptible than others to certain diseases. The prevalence of different *T*-cell classes varies among different individuals, and so their response to a superantigen would also vary. —*T.M.B.*

TECHNOLOGY

Bring Me Sunshine

Nonimaging optics brings full-strength sunlight to earth

Energy radiating from the sun is inexorably diluted as it disperses. By the time the solar radiation reaches the earth it is so weak that mad dogs and Englishmen, at least, can tolerate it with no more protection than that afforded by the atmosphere.

The dilution limits schemes to harvest solar energy. Many more applications would open up, engineers say, if there were a way to get hold of original-strength sunlight. Roland Winston and his colleagues at the University of Chicago have demonstrated how that can be done. They have built a collector that concentrates sunlight so that it reaches intensities that may exceed those at the surface of the sun.

Instead of resorting to conventional, image-making optics, whose ability to concentrate light is inherently limited, Winston has pursued a nonimaging approach. A parabolic mirror directs light into the wide end of a hollow silver cone capped with a lens and filled with highly refractive oil. The secret lies in the shape of the cone's interior, which follows a mathematically complicated curve. The curve, Winston explains, is shaped according to the "edge ray" method: all light beams from the most extreme intake angle are reflected no more than once off the cone's interior before reaching an output hole at the narrow end. "That," Winston says, "gives you the maximum possible concentration."

A tabletop Winston device can magnify sunlight 56,000-fold, achieving an irradiance of 4.4 kilowatts per square centimeter. By replacing the oil-filled silver cone with a solid cone of yttrium-aluminum-garnet (YAG, a favorite of laser makers), the workers hope to concentrate sunlight 80,000-fold: YAG refracts even more strongly than the oil, and it can reflect light internally with essentially no losses.

The laboratory demonstration of Winston's design, described in *Nature*, had an output beam just a millimeter in diameter. The Solar Energy Research Institute (SERI) laboratory in Golden, Colo., plans to build a Winston device with a primary mirror a meter or so in diameter; the output beam's diameter would span a few centimeters.

Other types of nonimaging solar collectors have already begun to find commercial applications in solar-power installations in Israel and Japan. The potential of Winston's invention for power generation is now being explored. Many other kinds of applications are also in prospect for the Winston device, according to John P. Thornton of SERI. Thornton suggests that extremely bright sunlight could be used in high-temperature materials processing, for destroying toxic chemicals and for pumping lasers. *—T.M.B.*

Giant Flux Creep

The new superconductors: raw spaghetti versus slime

When high-temperature superconductors were first popularized in 1986, some scientists and many reporters predicted an early millennium: superfast computers, levitating trains and high-strength magnets would transform the world economy. All these devices would be cooled by liquid nitrogen, a substance "less expensive than beer." Three years later, reporters are spreading gloom instead. "Hopes for Superconductivity Begin to Fade," the *New York Times* proclaims. "Is the Party Over?" *Science* asks.

Underlying the grim headlines is not some great anti-breakthrough but rather a suddenly publicized competition between two theories that purport to explain why high-temperature superconductors do not carry large currents easily when subjected to magnetic fields. This shortcoming puts cheap electromagnets for magnetic resonance imaging and the oft-imagined levitating train out of reach for the present; development of applications that do not involve intense currents or fields, such as superconducting magnetic-field probes and lossless conductors for computer chips, is progressing apace, researchers say.

Physicists and materials scientists have known for well over a year that passing a current through a high-temperature superconductor immersed in a magnetic field exerts a force on the lines of magnetic flux that penetrate the superconductor, pulling them along and creating electrical resistance. What workers are debating is whether the lattice of flux lines remains essentially solid while it deforms or undergoes a phase transition like that of ice melting into water. Experiments that might settle the issue have yet to be done.

One camp, which includes Alexis P. Malozemoff of the IBM Thomas J. Watson Research Center, opts for "giant flux creep." In this process individual lines jump from one spot to another under the force of a current; eventually the entire lattice slides downstream.

David J. Bishop of AT&T Bell Laboratories, whose electron microscope pictures of jittering flux lines gave visual interest to the story, opts for melting instead. David R. Nelson of Harvard University, another proponent of melting, likens the solid lattice to a boxful of uncooked spaghetti; as the temperature increases, he says, the flux lines soften and flow until finally they form a tangled mass.

In fact, Nelson's calculations predict that the flux lines could behave like long-chain polymers such as Silly Putty or Slime, reacting like solids to sudden forces but flowing slowly over longer periods. Even limp flux lines could strongly resist flow, he says.

Meanwhile, regardless of who is right, materials scientists are working to make superconductors whose flux lattices will remain stable. Superconducting thin films, which carry high currents in high magnetic fields, "constitute an existence proof" for structures that can hold a flux lattice in place, says Malcolm R. Beasley of Stanford University. The thin films appear to have more crystal defects that can trap and hold flux lines, according to Sungho Jin of Bell Labs, but no one knows how to make similar defects in bulk superconductors.

Jin's colleague Robert B. van Dover points out that irradiating yttrium-based bulk superconductors with fast neutrons can increase their critical currents by a factor of 10, which suggests the possibility of similar gains for materials with higher critical temperatures as well. But Bishop points out that the same trait giving some superconductors high critical temperatures—weak coupling between conducting planes—makes their flux lattices flexible and easy to melt.

Glaring publicity and the time taken by new discoveries to reach the market combine to make further swings between exultation and despair inevitable, Bishop suggests. Alan M. Wolsky of Argonne National Laboratory cautions: "People shouldn't be discouraged by the realization that the task at hand is difficult." *—Paul Wallich*

OVERVIEW

Fifth AIDS Conference

Is HIV spreading at epidemic rates in the inner cities?

MONTREAL—One had to admire the activists who seized the stage in the main hall of the Palais des Congrès and delayed the opening ceremonies of the Fifth International AIDS Conference by an hour. With their hand-lettered signs, bright T-shirts, catchy slogans and civil disobedience, they recalled the civil-rights movement or the movement against the war in Vietnam. But they had taken up these methods on behalf of a new cause: better and more equitable treatment for those who have AIDS. The pain of the protestors, many of whom are infected with the AIDS virus, was clearly real; their protest was, for the most part, restrained.

That was Sunday, June 4. By the end of the conference on the following Friday, the gay activists had come to seem a bit out of date, like the generals who are always preparing to fight the last war. As many presentations here indicated, the focus of the epidemic is no longer among homosexual men. It is among drug users and their sexual partners. What is more, preliminary evidence suggests HIV may already have begun to radiate outward from intravenous drug users in the inner city, spreading in conjunction with cocaine use, prostitution and an explosion of other sexually transmitted diseases whose presence increases the risk of transmitting HIV.

It has been apparent for some time that the AIDS virus is spreading more rapidly among intravenous drug users than among any other known risk group. Largely as the result of education and safer sexual practices, the infection rate among homosexual men seems to be slowing. Because of the long lag time between infection and disease, however, the number of cases of AIDS among intravenous drug users has remained lower than the number among homosexual men. Now that pattern, too, is changing, as Roel A. Coutinho of the Netherlands Department of Public Health and the Environment reported at the plenary session on Wednesday. Coutinho predicted that 1990 will be the first year in which more new cases of AIDS are diagnosed in Europe among drug users than among homosexual men.

The picture in the U.S. is more complex, Coutinho noted. From 1985 through 1987 the proportion of new AIDS cases accounted for by users of intravenous drugs remained stable at 19 percent; in 1988 it rose to 26 percent. Even this increase probably understates the impact of drug use on the epidemic in the U.S. Studies in New York City by several investigators have shown that intravenous drug users there who are infected with HIV often die of other diseases (including pneumonia or tuberculosis) before they develop full-blown AIDS.

One of the chief concerns about the epidemic of HIV infection among drug users is that they may form a conduit that carries the virus into other groups in the population. Some evidence presented at Montreal suggests that this has indeed begun to happen. For one thing, the proportion of all AIDS cases that results from heterosexual transmission has risen steadily from 1 percent in 1985 to about 5 percent in 1987. Many of these cases are found in sexual partners of intravenous drug users, but the epidemic has apparently not stopped there.

Data gathered by the Centers for Disease Control indicate that in some metropolitan areas the virus has begun to spread away from its initial focus in drug users. Michael E. St. Louis of the CDC presented results from a "sentinel" survey of 27 U.S. hospitals that began in January, 1989. Leftover clinical blood samples were tested for the AIDS virus after excluding all samples from people with medical conditions suggesting HIV infection. Infection rates varied greatly; in some regions they were as low as .1 percent. In the New York metropolitan area, however, two hospitals serving poor communities had rates of infection ranging from 5 to 7 percent.

In the inner-city areas where rates of infection are high, St. Louis said, relatively more women are infected, suggesting that heterosexual transmission is more significant there. Furthermore, he said, when the population is divided by age, HIV infection begins to appear in those as young as 12 years old. "HIV has reached epidemic proportion among the poor served by these inner-city hospitals," St. Louis said. Because of the great variation in rates of infection among communities, he added, "the burden of HIV infection will continue to fall unequally on those who can least afford it."

What epidemiological mechanisms are at work in the communities hit hardest by HIV infection? Few of the available studies can provide a direct answer. But work by King K. Holmes of the University of Washington and his colleagues provides disturbing insights into the epidemiology of sexually transmitted disease in poor, minority, urban neighborhoods.

In his address to the plenary session on Thursday morning, Holmes noted that in the early 1980's rates of sexually transmitted disease were falling among both blacks and whites in the U.S. Since then, he said, rates have continued to fall among whites but have risen sharply among blacks. For example, the decline in new cases of gonorrhea was reversed in the black population in 1985; the rate has continued to rise since then. As a result of these divergent trends, by 1988 the incidence of new cases of gonorrhea was 21 times higher in black women than it was in white women.

To supplement these national data, Holmes and his colleagues conducted a detailed survey of gonorrhea in King County, Wash., the county that includes Seattle and its suburban fringes. For each race and ethnic group the highest rates of gonorrhea by far were seen among 18- and 19-year-old women. Absolute rates were highest in the few census tracts where the indicators of socioeconomic status are lowest. In those neighborhoods, from 2 to 3 percent of the population are catching gonorrhea every year, constituting an aggravated epidemic.

Because the study included analysis of laboratory results, the appearance of a unique strain of gonococcus in Seattle in 1986 enabled Holmes's group to follow one microepidemic in detail. The first cases of infection with this strain were seen in white and Asian men. As the episode continued in 1987 and 1988, however, the strain became concentrated almost exclusively in blacks, who make up some 5 percent of the Seattle population.

As the strain spread, the epidemic changed, becoming entangled with illegal drug use and with exchanges of sex for money. At the beginning of 1987, 19 percent of those who had contracted the strain said they had recently used illegal drugs. By the end of 1987 the figure was 82 percent. The fraction of those infected who said they had engaged in commercial sex remained high throughout the epidemic. As a result, the proportion of infected people who acknowledged both illegal drug use and prostitution increased from 10 percent at the beginning of the epidemic to 71 percent in its late stages.

Much of the drug use reported in this study was attributed by Holmes to the spread of "crack," the smokable form of cocaine that began to reach epidemic proportions in some cities during 1985. Some female crack addicts trade sex for drugs, exacerbating the spread of gonorrhea and other diseases. In consequence, Holmes said, these sexually transmitted diseases are becoming concentrated in "core groups" consisting of young members of ethnic minorities among whom illegal drug use and commercial sex are common.

As the epidemic of sexually transmitted disease spreads among core groups, programs for treating such diseases are increasingly unable to meet the demand for services. Holmes cited a CDC study of 15 public clinics in the U.S. showing that most turn away more patients and make patients wait longer than they did only a few years ago. "For example," he said, "one large East Coast clinic in a city where syphilis is epidemic reported that in 1985 new patients were registered until 3 P.M. By 1985 they had to stop at 11 A.M. In 1985 patients waited on average two and a half hours before being seen by a doctor; in 1988 the wait was four hours."

"There is a clear need for a new emphasis on developing resources for diagnosis and treatment of sexually transmitted diseases in those clinics serving inner-city populations," Holmes continued. "And frankly I'm not sure the cities can afford to bear the full cost of this new emphasis." He concluded with a warning: "The last thing we can allow to happen today is continued epidemic spread of those sexually transmitted diseases that have been implicated as risk factors for sexual transmission of HIV in the very same inner-city populations that now have epidemic rates of HIV due to intravenous drug abuse and other factors." *—John Benditt*

SCIENTIFIC AMERICAN August 1989 Volume 261 Number 2

Alternatives to Animals in Toxicity Testing

The use of animals in evaluating chemical safety is costly, time-consuming and increasingly criticized by animal-welfare groups. Alternative methods can reduce the number of animals needed

by Alan M. Goldberg and John M. Frazier

Each year thousands of chemicals undergo rigorous testing designed to evaluate their potential toxicity. Almost all of the tests take place in animals: the reactions of rats, rabbits and mice to chemicals are currently the best available predictors of the effects the substances will have on the human organism. The introduction of animal testing in the U.S. in the 1920's was a major advance in toxicity testing, and subsequent debates about the place of animals in testing were qualified by the absence of better alternatives.

In the past decade the issue of whole-animal toxicity testing has become more urgent and contentious. Animal-welfare advocates have decried the suffering of millions of animals, and industries bringing chemicals to the marketplace have begun to chafe at the costs and delays imposed by animal testing. Meanwhile, case histories such as that of thalidomide serve to remind the public and testing establishments alike of the perils of letting unsafe chemicals reach the marketplace. In answer to these concerns, toxicologists began exploring possible alternatives.

Their exploration has yielded a new methodology known as in vitro toxicity testing. Literally, *in vitro* means "in glass," but biologists interpret the term more broadly to mean research that does not involve intact higher animals. In vitro testing includes a battery of living systems—bacteria, cultured animal cells, fertilized chicken eggs, frog embryos—that can be employed to evaluate the toxicity of chemicals in human beings. Ultimately workers hope to be able to test chemicals in cultures of human cells from various organs and tissues so that the question of human toxicity can be answered more directly.

Several factors have paved the way for the introduction of in vitro testing. One is the growth of the science of toxicology itself. Today investigators understand much better how toxicological processes are begun and how toxic effects are expressed; they need not use the death or illness of an animal as an end point in their studies. Another factor has to do with technological developments of the past few years. New options in culture techniques and bioanalytical tools allow workers to monitor toxicity with unprecedented thoroughness and precision at the cellular level rather than at the organismal level.

Yet the obstacles such efforts face are tremendous. Some are technical: singly or in combination, in vitro tests as yet cannot approximate the complexity of interactions that take place in a living animal. Some of the obstacles are bureaucratic: no framework has been established for approving in vitro test procedures or for incorporating the results of such tests into the evaluation of results from whole-animal, or in vivo, methods.

We believe in vitro tests will eventually be able to meet these challenges. Protocols for in vitro tests already exist that can complement the current panoply of whole-animal procedures and reduce the number of animals that are subjected to testing. It is not too soon to begin planning ways to integrate in vitro testing into toxicity testing as a whole.

Toxicity testing is one of the two major components of risk assessment, the process by which new substances are evaluated for their potential impact on human health and welfare. The other component is assessment of exposure. The exposure estimate indicates how many people

ALAN M. GOLDBERG and JOHN M. FRAZIER work together at the Center for Alternatives to Animal Testing at Johns Hopkins University. Goldberg, the center's director, served on the U.S. Congressional Office of Technology Assessment panel on alternatives to animals in testing, research and education and is currently a council member of the National Academy of Sciences' Institute for Laboratory Animal Resources. Frazier is director of the center's In Vitro Toxicology Laboratory and is currently involved in developing measures of cell death in cultures that would decrease the number of animals required for chemical testing.

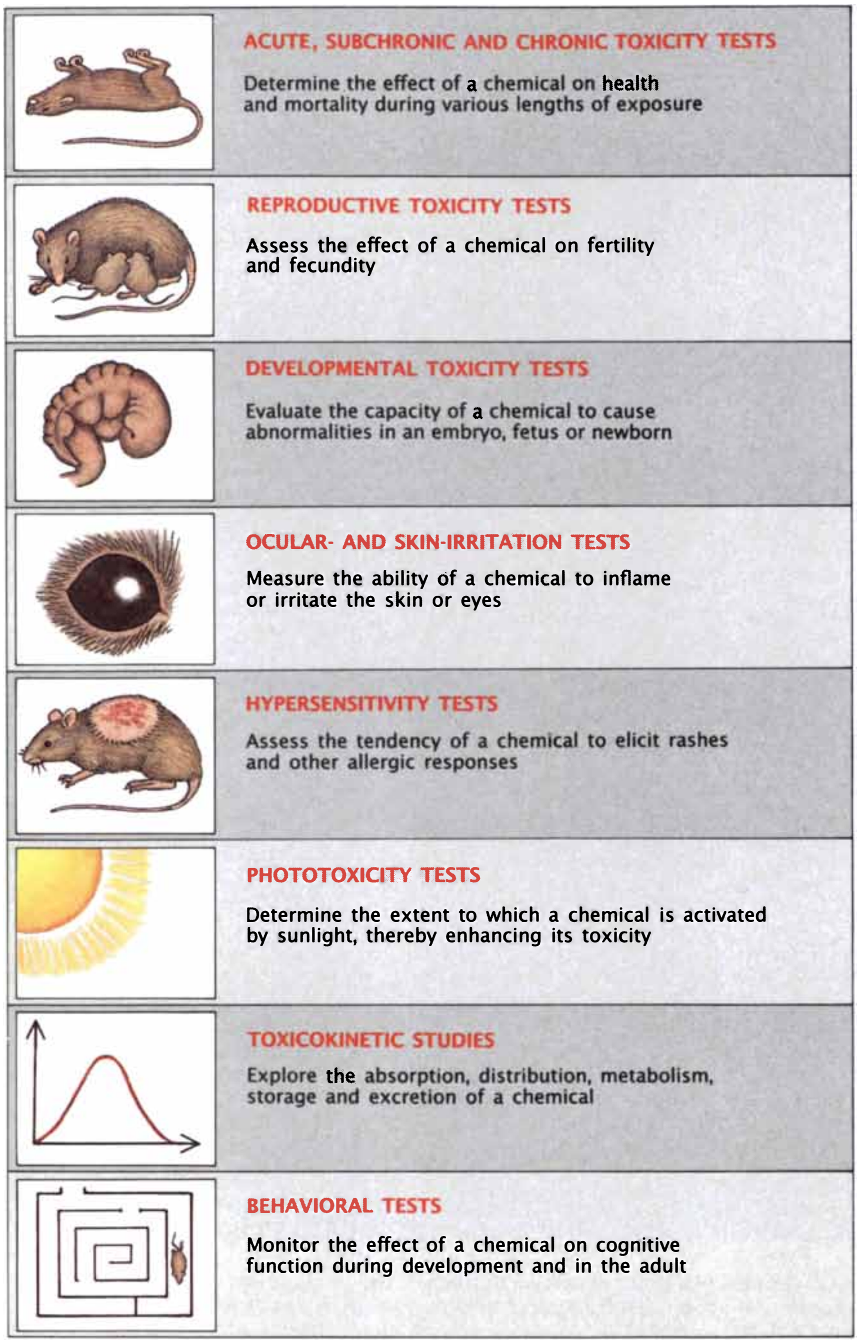
NEUTRAL-RED ASSAY helps to determine a chemical's effect on cultured cells. The cells and various concentrations of the test chemical are placed in the wells of the assay plate along with a dye called neutral red. Only living cells take up the dye, and so cultures of healthy cells turn pink (*left and middle of plate*), whereas cultures in which the cells have died are pale (*right*).

will be exposed to a chemical in what concentrations, for how long and under what conditions. For a chemical to pose a risk of notable proportion, there must exist the likelihood of human exposure to the agent in quantities sufficient to produce adverse biological effects.

Toxicity testing is required for new chemicals introduced into the marketplace, old chemicals that are proposed for new uses and new mixtures of old or new chemicals. The main objectives of such testing are twofold. The first objective, known as hazard identification, involves determining which potential adverse effects—cancer, kidney damage, reproductive injury and so on—can ensue from exposure to a given chemical. The second objective is to provide data estimating the quantitative exposure-response relationship for the chemical in human beings and other organisms [*see illustration on page 28*].

The exposure-response relationship describes the likelihood of an organism's developing a particular adverse biological response as a function of its exposure to the chemical. Such a relationship presumably exists for each of the hazards identified for a given chemical. It may vary, however, depending on how a person is exposed to the chemical—whether through ingestion, inhalation or contact with the skin. The age, genetic makeup and nutritional status of the person may play a role as well.

STANDARD WHOLE-ANIMAL TOXICITY TESTS



The LD50 test ("LD" stands for "lethal dose") is a classic example of an exposure-response test. A measure of acute lethality, the test was developed in the 1920's to determine the potency of digitalis and other medicinal preparations derived from biological materials. It provides a statistically accurate measure of the amount of a chemical that will produce 50 percent mortality in a population of animals, that is, the amount of the chemical that will kill half of them. Comparison of LD50 values for different agents gives a measure of relative toxicity. A variation on the LD50 test is the ED50 (for "effective dose"), which measures the amount of a chemical that will produce a deleterious effect other than death in 50 percent of the population.

The Draize ocular- and skin-irritation tests are other classic indexes that are widely used today. John H. Draize of the Food and Drug Administration standardized the protocol for the ocular-irritation test in the 1940's; it dictates specific procedures for measuring eye irritation in rabbits. A fixed dose of a chemical (.1 milliliter of a liquid or .1 gram of a solid) is placed in one of a rabbit's eyes; the other eye serves as a control. For the skin test, an area of the rabbit's hide is shaved and covered with the chemical being tested. In both tests there is a specific set of criteria for scoring irritation and inflammation.

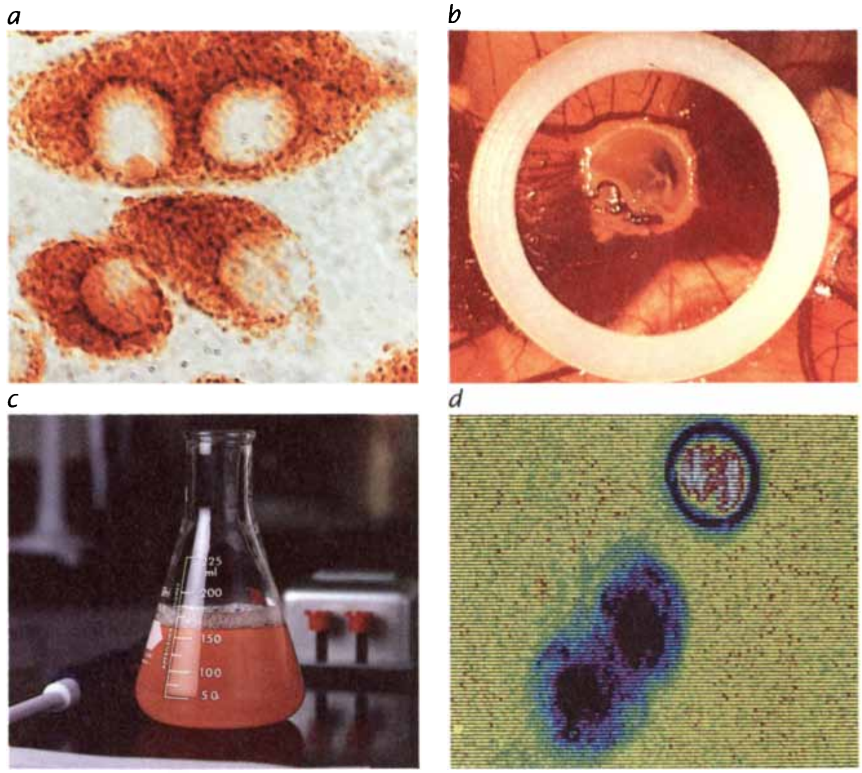
The LD50 test and the Draize tests are probably the toxicity tests most familiar to the public; they are also the ones singled out most frequently by animal-welfare activists. But chemicals are usually screened through many additional in vivo tests. These include acute toxicity tests other than the LD50 and subchronic and chronic toxicity tests that last anywhere from two weeks to two years. Such tests provide information on mechanisms of action, target organs, symptomatology and carcinogenicity (the ability to cause cancer) as well as lethality.

Other tests help to fill out the toxic profile of a chemical. Reproductive and developmental toxicity tests evaluate chemicals' effects on reproductive success and their ability to cause developmental malformations, a property known as teratogenicity. Hypersensitivity procedures test for chemicals that may not directly damage the skin but may elicit instead an immunological response similar to the one produced by poison ivy. Phototoxicity testing determines whether sunlight will activate a test chemical and thus enhance skin irritation.

Studies of toxicokinetics are sometimes carried out to trace the absorption, distribution, metabolism, storage and excretion of a chemical. Such studies are quite useful when the same chemical exhibits differences in toxicokinetics in two animal species. Finally, neurological and behavioral tests monitor the effects of chemicals on cognitive functions in adult animals as well as in developing fetuses.

It is probably clear from this recitation of procedures that complete toxicological evaluation of even one chemical is complicated, time-consuming and expensive. Testing a typical new chemical costs between $500,000 and $1.5 million, takes up to two or three years and may entail the sacrifice of thousands of animals. Furthermore, tens of thousands of products already on the market have never been tested thoroughly. The National Academy of Sciences observed recently that many of those substances might not have been evaluated at all.

Obviously, enormous benefits would accrue from toxicity tests that are cheaper and faster than in vivo testing. Just as obviously, researchers will be hard-pressed to come up with a battery of in vitro tests that can match the exhaustive screening possible with whole-animal procedures. Workers are making progress on roughly half a dozen categories of techniques.



IN VITRO TOXICITY TESTS are tests that do not take place in whole animals. Several are in development or already in use. The neutral-red assay shown on page 25, for example, can be applied to cultures of human skin cells to provide information on inflammation and irritation (*a*). The membrane lining the shell of a fertilized chicken egg can serve the same function; here a chemical contained within the white Teflon ring has produced a mild lesion (*b*). Cultures of liver cells (*c*) can give indications of substance metabolism as well as of liver damage. False-color computer analyses (*d*) help toxicologists to interpret data gathered from in vitro results. Here the dark blotches represent dead cells; the red and white one corresponds to a live cell.

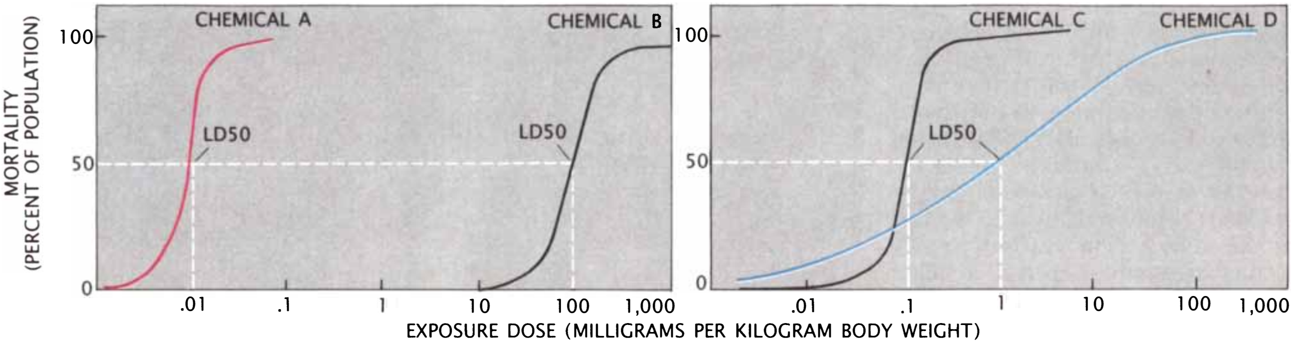
The area of in vitro testing that has been pursued the longest and has been the best funded is that of genotoxicity: the ability of a chemical to damage genetic material. Genotoxicity encompasses substances that cause cancer, gene mutation and chromosomal abnormalities. Whole-animal tests for carcinogenicity are among the most expensive and time-consuming toxicity tests, which is probably why more than $70 million has been spent in the U.S. over the past decade to find in vitro alternatives. The in vitro tests currently available, such as the standard Ames bacterial assay, are widely used to screen for potential genotoxicity, but they cannot be expected to preempt whole-animal tests of chronic exposure, such as the rodent lifetime bioassay.

Another area of in vitro testing that has a relatively long history is cytotoxicity testing. Simply put, cytotoxicity assays evaluate the ability of a substance to kill cells. Some of these assays were developed for special purposes, such as screening drugs for the ability to kill cancer cells; others are meant for more general use. The number of methods available for distinguishing dead cells from live cells has increased rapidly in the past several years. In fact, the limiting factor on cytotoxicity testing is the number of cell types that can be cultured in vitro.

Two test systems have received considerable attention in the context of in vitro cytotoxicity testing: the total cellular-protein assay and the neutral-red uptake assay. In both tests cells cultured in plastic petri dishes are treated with various concentrations of a test chemical added to the culture medium. After 24 hours of exposure, the test chemical is washed out of the medium and an analytical reagent is added. In the case of the total cellular-protein test, a reagent called kenacid blue is added to the medium and reacts with proteins in the cells, imparting a blue color whose density can be measured. Healthy, rapidly growing cells contain more protein than dead ones; consequently, control cultures will be dark blue. Dishes in which cells have been killed by the test chemical will be progressively lighter in tone.

The concentration of the test chemical that produces 50 percent inhibition of protein content, known as the IC50, can be determined from the colors of the cell cultures and compared with the IC50's of known chemical toxins in order to rank the test chemical's relative toxicity. This assay can be automated to speed testing, and it can be performed in combination with enzymes that metabolize drugs so that the effects of chemical intermediates can be tested as well.

The assay measuring neutral-red uptake is not too different. Developed in its present form at Rockefeller University, the test is based on neutral red,



DOSE-RESPONSE CURVES help to determine the concentration of a chemical that causes death in 50 percent of a population, a figure known as the LD50 (*left*). The LD50 can serve as a useful index of toxicity; toxicologists plot the entire curve, however, because at low concentrations a chemical with a high LD50 can sometimes be more toxic than one with a low LD50 (*right*).

a dye that is taken up from the culture medium and stored by living cells but not by dead ones. The amount of dye retained by the cells is an indication of the number of living cells. Again, an IC50 for the test chemical is established by linking cell mortality to the amount of the chemical the cells received. The assay is then quantified by comparing the IC50 with the IC50's of known toxins.

Although they may have only limited ability to predict tissue-specific effects or effects resulting from tissue or organ interactions, cytotoxicity tests do provide essential information on the intrinsic toxicity of pure chemicals, mixtures and formulations. They can also be fairly good indicators of ocular irritation, because cell death is a major cause of it. Corneal epithelial cells can be subjected to cytotoxicity testing for ocular irritation, and in fact, such tests are already used in product safety evaluation.

The Center for Alternatives to Animal Testing at Johns Hopkins University has identified more than 30 other in vitro tests that could be appropriate for testing ocular irritation. Some of these also test for cytotoxicity, but others have different end points. Ray Tchao of the Philadelphia College of Pharmacy and Science, for example, has devised a protocol to detect impairment of the so-called tight junctions between cells, junctions that are important in controlling the penetration of substances through the corneal epithelial cell layer.

The so-called CAM test, pioneered by Joseph Leighton of the Medical College of Pennsylvania and by Niels P. Lupke of the University of Münster, provides another in vitro measure of inflammation. In the CAM test, part of the shell of a fertilized chicken egg is carefully removed to reveal the delicate, veined chorioallantoic membrane (CAM) underneath. A test chemical is applied directly to the membrane; sometimes a Teflon ring is also placed on the membrane to contain the chemical. Researchers look for inflammation of the membrane five minutes and 24 hours after the chemical has been applied.

Several laboratories have also been exploring cultures of human epidermal cells as models of human skin. Some skin-cell culture methods are descended from skin-regeneration techniques developed for burn patients. The skin-cell cultures can be tested for inflammation much as the membrane in a CAM test is. Measuring the biological response to chemicals, however, is easier in skin-cell cultures than it is in CAM tests.

In vitro tests are also being developed to monitor toxicity in particular target organs. The question of target-organ toxicity is answered in vivo by examining the organs of a treated animal for pathological changes. In vitro, cells from specific organs must be cultured and tested. Considerable progress has been made on in vitro screening for liver, blood, heart, kidney, lung and nervous-system toxicity.

The techniques for culturing hepatocytes (liver cells) are particularly well developed. Methods of in vitro hepatotoxicity testing, derived from experiments in liver research, involve isolated liver cells, liver slices and isolated, perfused whole livers. Human hepatocytes have already been used in some tests; where the cells of other animals are used instead, in vitro target-organ data can still reduce substantially the number of animals needed for conclusive results. Enough tissue can be obtained from two or three animals to conduct studies that would ordinarily require from 20 to 40 animals.

The purpose of these in vitro systems ranges from identifying chemicals that specifically produce toxicity in the liver to determining the metabolic kinetics of chemicals and the way in which they are excreted. Test systems based on rat hepatocytes can also evaluate cellular markers for potential toxicity. As more knowledge is obtained about the mechanisms of toxic action of chemicals in organ systems, new in vitro methods can be developed to test for these effects.

Progress has also been made in identifying in vitro systems for evaluating teratogenicity. The key to teratogenicity testing lies in establishing the relation between in vitro indexes of toxicological response and the complex process of differential toxicity in the developing organism, particularly in the human fetus. Although many of the alternative test systems proposed involve whole organisms—from the hydra or fruit fly to frog or rodent embryos—and therefore circumvent the problem of extrapolation between cell cultures and whole animals, the systems still have significant problems in predicting human teratogenicity.

In some cases, mathematical and computer models may be able to supplement the information provided by in vitro testing. Mathematical pharmacokinetic models are already helping toxicologists to estimate in vivo toxicokinetics from in vitro data. Computer-based "structure-activity" analyses attempt to correlate general toxicological responses with aspects of the molecular structure of the test chemical. Such methods are currently empirical, but they should improve as specific mechanisms of interaction are related to chemical structure.

The complexity of the exposure-response relationship seems, however, to rule out the possibility of making sound predictions from theoretical principles alone. The response of an organism to a given exposure of a chemical results from a diffuse array of interdependent processes at the molecular, cellular and organismal levels.

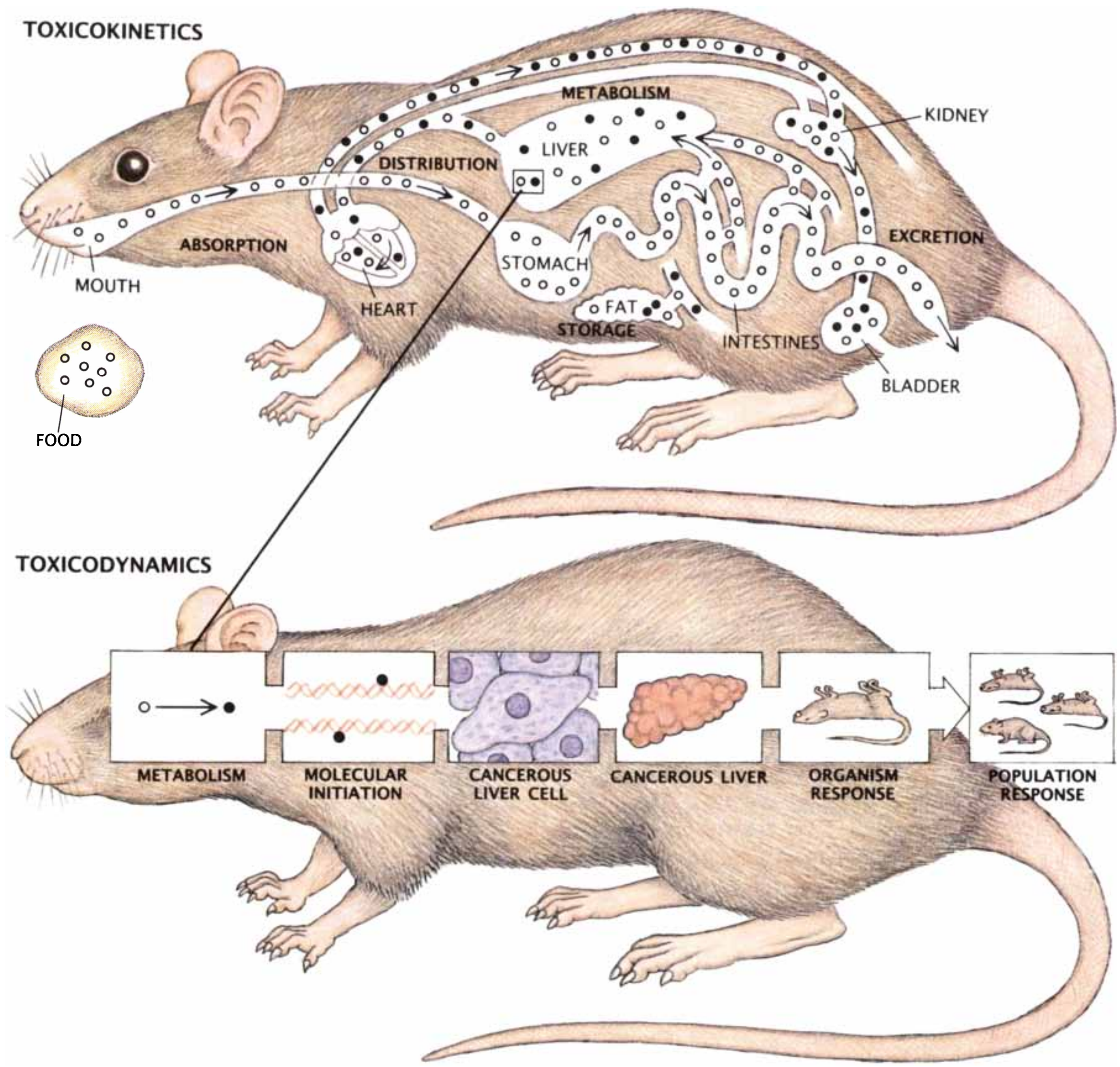
In addition to the absorption, distribution, metabolism, storage and excretion of substances described by toxicokinetics, the outcome of an organism's exposure to a chemical also depends on toxicodynamics.

Toxicodynamics has to do with alterations in the biological system that are a consequence of the presence of a chemical in the system [*see illustration below*]. At the molecular level such alterations are biochemical: they can, for example, inhibit an enzyme critical to normal cellular function. At higher levels of organization the alterations are manifested as tissue pathology or as clinical toxicity.

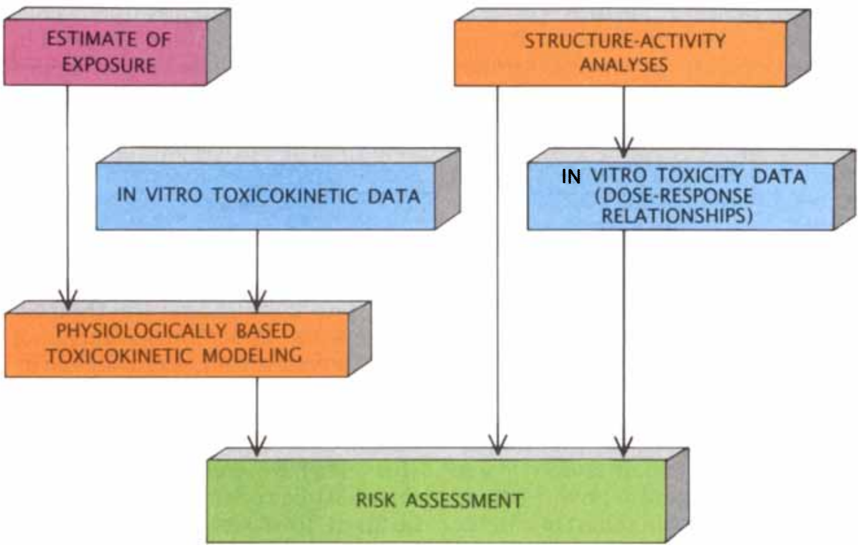
If a human being is exposed to a chemical, the toxicokinetic properties of the chemical determine whether the agent or one of its metabolites will ultimately reach a sensitive cellular or molecular target and initiate a biological response. If the reactive form of the chemical does reach the potential target, the toxicodynamics determine to what degree the agent will adversely affect the human being. The ultimate expression of pathology depends on the human organism's ability to repair toxin-induced damage at all levels of biological organization: the molecular and cellular levels as well as the levels of organs and tissues.

The toxicokinetics and toxicodynamics of a chemical and the ability of a biological system to effect repair all come to bear on the exposure-response relationship. It is easy to see why it is challenging to predict human toxicity with anything short of a whole organism. Indeed, current predictive toxicology draws not just from whole-animal experiments and theory but from a historical data base, compiled over years of experience, that relates the results of in vivo testing to human epidemiological data and even in some cases to the outcome of accidental human exposure.

The foregoing discussion of toxicokinetic and toxicodynamic interactions underscores the most obvious and important advantage of whole-animal testing: it provides an integrated biological system that serves as a surrogate for the complexities of human and other animal systems. In vivo tests have several other features that must be reckoned with. They can be used to assess the outcome of exposure by different routes (whether



TOXICOKINETICS AND TOXICODYNAMICS describe a chemical's interaction with a living organism. The chemical depicted here causes liver cancer in rats. The absorption, distribution, metabolism, storage and excretion of the chemical are specified by its toxicokinetics (*top*); the effects the chemical and its metabolites have on an organism are the toxicodynamics (*bottom*). Because they involve many levels and systems, toxicokinetics and toxicodynamics are hard to predict without animal testing.



HYPOTHETICAL RISK-ANALYSIS PROCEDURE would eliminate animals in toxicity testing and incorporate in vitro data (*blue*) and computer-based evaluations (*orange*). Although the realization of such a scheme is still many years away, some in vitro methods can be introduced now to reduce the number of animals needed in testing.

through ingestion, contact with the skin or inhalation) and over long periods (chronic toxicity tests can take a year or longer). In addition, whole-animal studies can be designed to determine whether or not particular toxic effects are reversible—an important parameter in risk assessment and risk management.

How might in vitro testing provide the same information? Would it be necessary to have one in vitro test for every potential target-cell type in the body? How can in vitro tests evaluate toxicological responses that involve, say, immunological processes or blood pressure? How can they evaluate chronic toxicity or recovery from toxic insults? How can exposure by ingestion, inhalation or topical contact be simulated? These problems must be solved if in vitro tests are ever to replace in vivo testing completely.

Another potential stumbling block concerns the testing of human-cell cultures, which, because it would eliminate the need for species extrapolation, is billed as one of the benefits of in vitro testing. There are a few hitches. Currently not all human cell types can be cultured; some types of cells "dedifferentiate" in vitro, that is, they take on the qualities of primitive, unspecialized cells instead of retaining the characteristics that identify them as muscle cells, spleen cells, colon cells and so on. Furthermore, the supply of normal human cells available for toxicological testing is somewhat limited. In order for human cells to be routinely employed in toxicity testing, some means of making them more readily available must be found.

These obstacles should be weighed against the disadvantages of whole-animal testing that we have already mentioned: animal discomfort and death, species-extrapolation problems and excessive time and expense. In vitro tests could ameliorate all of these problems and several more. For example, whereas whole-animal testing is hard to standardize, the standardization of in vitro techniques is fairly straightforward. Furthermore, the dose of a chemical that is received by each cultured cell can be measured and controlled with precision, making it easier to establish the critical concentrations of toxins. Because much smaller quantities of a substance can be used, novel compounds available in limited amounts can be tested, and disposal problems are minimized if a compound turns out to be toxic.

It takes time to overcome the problems inherent in introducing any new technology. It also takes time to gain acceptance for a new technology when the incumbent technology can boast a 50- to 60-year record of empirical findings. Before any new in vitro test can become a regular, routine source of toxicological data, it will have to be validated. That means it must be shown to be reliable (to give consistent results in different laboratories and at different times in the same laboratory) and meaningful (to provide information that contributes to chemical safety evaluation). To promote acceptance, toxicologists must also begin to compile data bases for in vitro tests so that better predictions can be drawn from the results.

Contrary to a prevailing misperception, in vitro tests need not replace existing in vivo test procedures in order to be useful. They can contribute to chemical safety evaluation right now. In vitro tests, for example, can be incorporated into the earliest stages of the risk-assessment process; they can be used to identify chemicals having the lowest probability of toxicity so that animals need be exposed only to less noxious chemicals. Such a procedure would reduce the number of animals tested and would also save time and the expense of research and development for products likely to fail subsequent safety evaluations. It is encouraging to note that several corporations have already implemented this approach in their testing strategies.

In any case, insisting on comprehensive replacement of existing tests will only delay the implementation of in vitro methods indefinitely. In vitro toxicity testing will not replace animal testing in a single, quantum step. In fact, regulatory mechanisms do not yet exist for review and approval of new in vitro testing methodologies for chemical safety evaluation. With time, in vitro testing will become more firmly established, and it will eventually play a critical role in the safety-evaluation process. It is our hope that this goal will be attained with the support and encouragement of industry, regulatory agencies, the scientific community and animal-welfare advocates alike.

FURTHER READING

IN-VITRO METHODS MAY OFFER ALTERNATIVES TO ANIMAL TESTING. Ron Dagani in *Chemical and Engineering News*, Vol. 62, No. 46, pages 25–28; November 12, 1984.

A CRITICAL EVALUATION OF ALTERNATIVES TO ACUTE OCULAR IRRITATION TESTING. John M. Frazier, Shayne C. Gad, Alan M. Goldberg and James P. McCulley. Mary Ann Liebert, Inc., 1987.

PHARMACOKINETICS IN RISK ASSESSMENT: DRINKING WATER AND HEALTH, Vol. 8. National Research Council. National Academy Press, 1987.

PREDICTION OF CHEMICAL CARCINOGENICITY IN RODENTS FROM IN VITRO GENETIC TOXICITY ASSAYS. Raymond W. Tennant et al. in *Science*, Vol. 236, No. 4804, pages 933–941; May 22, 1987.

FOURTH INTERNATIONAL WORKSHOP ON *IN VITRO* TOXICOLOGY. Edited by M. Balls and L. J. King in *Xenobiotica: The Fate of Foreign Compounds in Biological Systems*, Vol. 18, No. 6; June, 1988.


You're looking at a 2500% blowup with detail never before possible in a 35mm color print film.
Kodak introduces EKTAR color print film for the advanced photographer.
EKTAR 25 film delivers microfine grain for the most superior image structure; the highest resolution; the sharpest detail ever achieved in color print film.
Detail so sharp, you can even read the numbers on the seats.
New EKTAR film.
The genius is in the details.
Photograph taken on EKTAR 25 ISO film f/11 at 1/60sec. Available in 25, 125, 1000 ISO. For more information on new EKTAR film, call 1-800-242-2424.
Kodak
EKTAR 25
ULTRA SHARP COLOR PRINT FILM

The Great Supernova of 1987

On February 23 of that year astronomers gained their first closeup view of a star's cataclysmic death since 1604. Worldwide observations have tested existing theory and added new puzzles

by Stan Woosley and Tom Weaver

The collapse and explosion of a massive star is one of nature's grandest spectacles. For sheer power nothing can match it. During the supernova's first 10 seconds, as the star's core implodes to form a neutron star, it radiates as much energy from a central region 20 miles across as all the other stars and galaxies in the rest of the visible universe combined. To put it another way, the energy of that 10-second burst is 100 times more than the sun will radiate in its entire 10-billion-year lifetime. It is a feat that stretches even the well-stretched minds of astronomers.

Yet supernovas are more than distant spectacles: they make and expel the seeds of life. Only the simplest and lightest elements, hydrogen and helium, were formed in the primordial fireball of the big bang. Most of the heavier elements, including the carbon of our chemistry, the iron in our blood and the oxygen we breathe, were forged in supernovas long before the solar system took shape.

Important as they are, few supernovas have been seen nearby. The last one in our own galaxy flared in 1604, shortly before the invention of the telescope; Johannes Kepler, who observed it, was able to record only its brightness and duration. In the absence of nearby events, understanding of many features of supernovas has remained largely theoretical. Telescopes do reveal a dozen or so events each year in distant galaxies, and careful study of a few distant supernovas has served for testing some coarser aspects of theory. But none was close enough for the modern panoply of ground- and space-based instruments to chronicle the event in detail.

All that changed on the night of February 23, 1987, when a burst of light and a pulse of the elusive particles called neutrinos reached the earth from the brightest supernova in 383 years. Light from the explosion, 160,000 light-years away in the Large Magellanic Cloud, a satellite galaxy of our own, was visible only in the Southern Hemisphere. It is a tribute to the care with which amateur and professional observers monitor the southern sky that the supernova was photographed within an hour of the time its first light must have arrived—although the observer, Robert McNaught of Siding Spring, Australia, did not realize he had captured it until later.

About 20 hours after McNaught's first photograph, Ian Shelton of the Las Campanas Observatory in Chile was photographing the Large Magellanic Cloud. Comparing a photograph made that night with one from the night before, he found a new, starlike image on the later plate. The image was very bright—so bright that it ought to be visible to the naked eye. Shelton walked outside and looked up. Supernova 1987A (*A* for the first supernova, bright or faint, to be found that year) had been discovered.

Within a day anyone who had any astronomical instrument in the Southern Hemisphere was marveling at the sight. During the following months the array of instruments trained on the supernova came to include telescopes and sensors on board balloons, rockets, satellites and an airplane, as well as ground-based instruments of all descriptions. By now, more than two years later, the supernova has been studied at all wavelengths of the electromagnetic spectrum, and it is the first astronomical source of neutrinos to have been detected other than the sun. Together the observations give a coherent picture of the grand event, a picture that vindicates theory but also holds some surprises.

A supernova's characteristics are shaped by the progenitor star. In the broadest terms, SN 1987A is a type II supernova, powered by the gravitational collapse of a stellar core—a catastrophe unique to massive stars. (Type I supernovas, which include the 1604 event, are thought to be thermonuclear explosions of white-dwarf stars to which a critical mass of material has been added.) To make sense of what was observed in SN 1987A, it is best to begin with the history of the star that exploded. The story that follows is based on computer simulations of the evolution of a hypothetical massive star, which we and others (including Ken'ichi Nomoto and his colleagues at the University of Tokyo and W. David Arnett of the University of Arizona) have developed over the past 25 years in an effort to understand type II events. Since the supernova—the first to occur in an identified star—we have recalculated our model to take into account the special features of the star known beforehand as Sanduleak –69° 202, after the astronomer Nicholas Sanduleak, who catalogued it about 20 years ago.

The story begins about 11 million years ago in a gas-rich region of the Large Magellanic Cloud known as 30 Doradus, or the Tarantula Nebula, where a star was born with about 18 times the mass of the sun. For the next 10 million years this star, like most others, generated energy by fusing hydrogen into helium. Because of its

STAN WOOSLEY and TOM WEAVER have collaborated for more than 15 years in trying to understand the evolution and explosion of stars and the origin of the chemical elements. Woosley is professor of astronomy and astrophysics at the University of California, Santa Cruz, and chair of the department; he is also a member of the General Studies Group in the physics department of the Lawrence Livermore National Laboratory. He professes a general interest in all things that explode. Weaver heads the General Studies Group at Livermore and is chief scientist of the X-ray laser program. In addition to astrophysics and short-wavelength lasers, he studies controlled fusion and nonlinear dynamics. Weaver's other activities include yoga, aikido and cross-country running.

great mass the star had to maintain high temperatures and pressures in its core to avoid collapse; as a result it was much more luminous than the sun—about 40,000 times as bright—and a profligate burner of nuclear fuel.

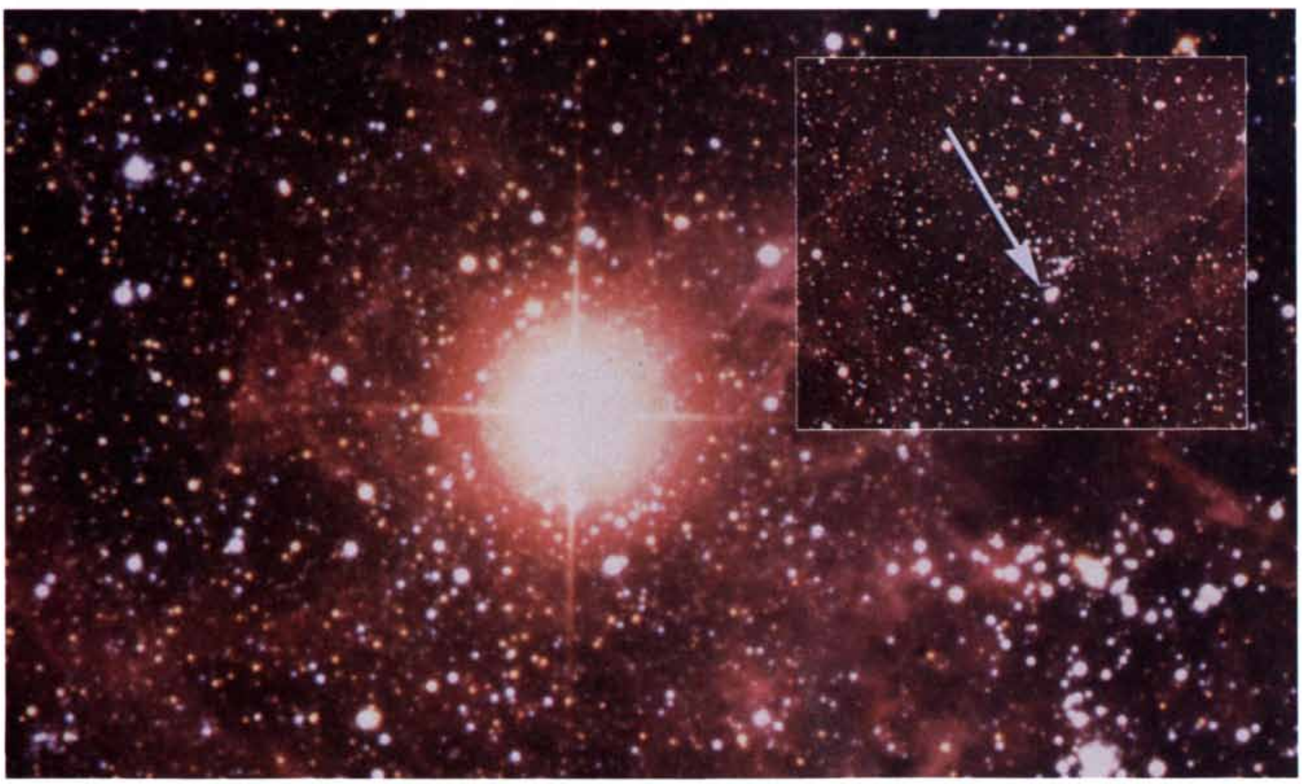
When hydrogen had finished fusing into helium in the innermost 30 percent of the star, the central regions began a gradual contraction. As the core was compressed over tens of thousands of years, from a density of six grams per cubic centimeter to 1,100, it heated up from about 40 million degrees Kelvin to 190 million degrees. The higher core temperature and pressure ignited a new and heavier nuclear fuel, helium. At the same time the outer layers of the star (mostly unburned hydrogen) responded to the additional radiation from the hotter core by expanding to a radius of about 300 million kilometers, or about twice the distance from here to the sun. The star had become a red supergiant.

The core's supply of helium was exhausted in less than a million years, burned to carbon and oxygen. During the few thousand years that remained to the star, this scenario of core contraction, heating and ignition of a new and heavier nuclear fuel—the ash of a previous cycle of fusion—was played out repeatedly. Carbon was the next to burn, at a core temperature of 740 million degrees K and a density of 240,000 grams per cubic centimeter, yielding a mixture of neon, magnesium and sodium. Then came neon, at 1.6 billion degrees and 7.4 million grams per cubic centimeter, followed by oxygen (2.1 billion degrees and 16 million grams per cubic centimeter) and finally silicon and sulfur (3.4 billion degrees and 50 million grams per cubic centimeter). Because ignition of successively heavier fuels took place in the very center of the star while previous fuels continued to burn in the less dense, overlying regions, the interior of the star came to resemble a cosmic onion, with elements layered in order of increasing atomic weight toward the center.

The core of the star passed through consecutive stages of burning at an accelerating pace. Whereas the burning of helium had lasted nearly a million years, carbon took 12,000 years, neon perhaps 12 years, oxygen four years and silicon, at the end, just a week. Each stock of nuclear fuel after hydrogen released about the same total energy, but at core temperatures above 500 million degrees K, beginning with carbon-burning, the star found a new and far more efficient way to spend its energy capital. Very energetic gamma-ray photons, abundant at such temperatures, were transformed into particle pairs—an electron and a positron, an electron's antimatter counterpart—as they passed near atomic nuclei. The particles promptly annihilated each other, usually re-creating the gamma rays but sometimes giving rise to neutrinos.

Neutrinos hardly interact with matter at all. They escaped from the star far more easily than the original gamma rays could have, carrying off energy. Even during carbon-burning, neutrino energy loss exceeded the energy loss by radiation. As the core's temperature rose during the later stages of its evolution, the neutrino luminosity rose exponentially to become a ruinous energy drain, hastening the star's demise.

This late evolution of the core proceeded too fast to have any effect on the star's vast envelope of hydrogen. Yet it turned out that the envelope had also evolved since the star had become a red supergiant.



AGED STAR AND ITS BRILLIANT DEATH are seen in photographs of the same region of the Large Magellanic Cloud made a few months apart. The progenitor star (*inset*), a blue supergiant called Sanduleak –69° 202, was about 80,000 times brighter than the sun; at its brightest (in May, 1987), the supernova reached 200 million solar luminosities. Even so, light represented only a tiny fraction of the total output of supernova 1987A: 30,000 times more energy was discharged in a burst of elusive particles called neutrinos. The photographs were provided by David F. Malin of the Anglo-Australian Observatory.

When workers first determined which star had exploded, they were surprised to find the progenitor star was not a red supergiant, as most stellar-evolution models for type II supernovas had predicted, but a blue supergiant—a smaller and hotter star [see "Helium-rich Supernovas," by J. Craig Wheeler and Robert P. Harkness; SCIENTIFIC AMERICAN, November, 1987].

The star's envelope, and not just its core, had apparently contracted beginning perhaps 40,000 years before the explosion, after the helium that had powered its red-supergiant stage was exhausted. Theorists are still debating the reasons, but the distinctive composition of star-forming gas in the Large Magellanic Cloud may have been the most important factor: in comparison with our own galaxy, the gas has a much lower content of elements heavier than helium. Among those elements, oxygen plays a special role in the evolution of a star. A lower oxygen content makes a star's envelope more transparent to radiation and hence perhaps more likely to contract. Oxygen also serves as a catalyst in the generation of energy by hydrogen fusion. Modeling suggests that a low initial oxygen content might subtly modify the early evolution of a massive star so as to ultimately yield a blue, rather than a red, supergiant.

The small radius of the progenitor star was to have dramatic effects later, when the star exploded, but it was irrelevant to the drama about to take place in the core. The week-long fury of silicon- and sulfur-burning had left the star with a core of iron, together with other elements in the iron group: nickel, chromium, titanium, vanadium, cobalt and manganese. Vast neutrino losses continued unabated because of the high core temperature, but having reached iron, the core had no nuclear currency left to pay its energy debt. Iron lies at the bottom of the curve of binding energy: energy must be added to fuse it into heavier elements or to split it into lighter ones. Fusion could go no further, and temperature and pressure could no longer maintain the core's equilibrium. Gravity won the 11-million-year contest, and the core began to collapse.

As the core was compressed, it did get hotter but not hot enough to stop the collapse. Two instabilities (discussed by William A. Fowler of the California Institute of Technology and Fred Hoyle, then of Cambridge University, during their pioneering theoretical work on supernovas in the early 1960's) actually accelerated the collapse. In one process, photodisintegration, high-energy photons tore apart the iron nuclei into lighter components, mainly helium—in effect reversing the fusion reactions of the star's previous history. In the second process, electron capture, free electrons were squeezed into nuclei, where they combined with protons to form neutron-rich isotopes. Both processes consumed energy, sapping critical support from the core; electron capture also removed some of the free electrons that had been a major source of pressure.

In a few tenths of a second the iron core, 1.4 times the mass of the sun and half the size of the earth, collapsed into a ball of nuclear matter about 100 kilometers in radius. When the center of the incipient neutron star exceeded the density of an atomic nucleus—270 trillion grams per cubic centimeter—the inner 40 percent of the core rebounded as a unit. The outer core, still plunging inward at close to a quarter of the speed of light, smashed into the rebounding inner core and rebounded in turn. A shock wave was born. In about a hundredth of a second, it raced out through the infalling matter to the edge of the core [see "How a Supernova Explodes," by Hans A. Bethe and Gerald Brown; SCIENTIFIC AMERICAN, May, 1985].

Workers modeling supernovas had hoped for many years that such a shock would continue outward through the many layers of the star, heating it and blowing it apart. Unfortunately, the most recent calculations for a star the size of Sk –69° 202, done by a number of theorists (including Sidney Bludman and Eric Myra of the University of Pennsylvania, Stephen Bruenn of the Florida Atlantic University, Edward A. Baron of the State University of New York at Stony Brook and Ron Mayle and James R. Wilson of the Lawrence Livermore National Laboratory), suggest that in SN 1987A the shock did not make it out of the core on its own.

The shock started out carrying enormous energy—about 10 times as much as was finally imparted to the exploding debris—but lost most of it beating outward against the infalling material. Photodisintegration and neutrino emission cooled the shock-heated material, sapping the shock's impetus. By the time the shock arrived at the edge of the iron core, the material behind it had no net outward velocity. The shock stalled and became an accretion shock, one through which material continuously flows inward. If this dismal state had persisted, the core would have swallowed the entire star. The result would have been a black hole, not a supernova.

Neutrino emission played a role in stalling the shock, and neutrino emission may also have helped to revive it. The core, having shrunk to a radius of 100 kilometers, had not reached nuclear density except at the center. It would become a true neutron star only when it had contracted to a radius of about 10 kilometers. Yet the protoneutron star was already very hot (Wilson and other modelers had predicted a temperature of about 100 billion degrees K) because of the gravitational energy released in the collapse. To contract further, the neutron star had to lose heat.

It did so through vast neutrino losses. The neutrinos were produced, as before, by the annihilation of electron-positron pairs made by the energetic gamma rays that pervade material at such high temperatures. This time, however, the neutrinos did not stream promptly out of the material: the density of the collapsing core was so high that it impeded even neutrinos. They diffused out of the core gradually, in seconds rather than milliseconds, slowing the star's contraction.

Even so, the power radiating from the contracting neutron star was outrageous, exceeding that of the rest of the visible universe. The total energy emitted in the 10-second neutrino burst was 200 or 300 times the energy of the supernova's material explosion and 30,000 times the energy of its total light output. It is now widely (but by no means universally) believed that a small fraction of the neutrino energy was somehow harnessed to revive the stalled shock and power the explosion. Building on a basic idea put forward in the mid-1960's by Stirling Colgate, now of the Los Alamos National Laboratory, Mayle and Wilson recently did a set of calculations that show just such an effect. Only a few percent of the neutrinos, interacting with the material just behind the stalled shock for about a second, deposit enough energy to accelerate the shock outward.

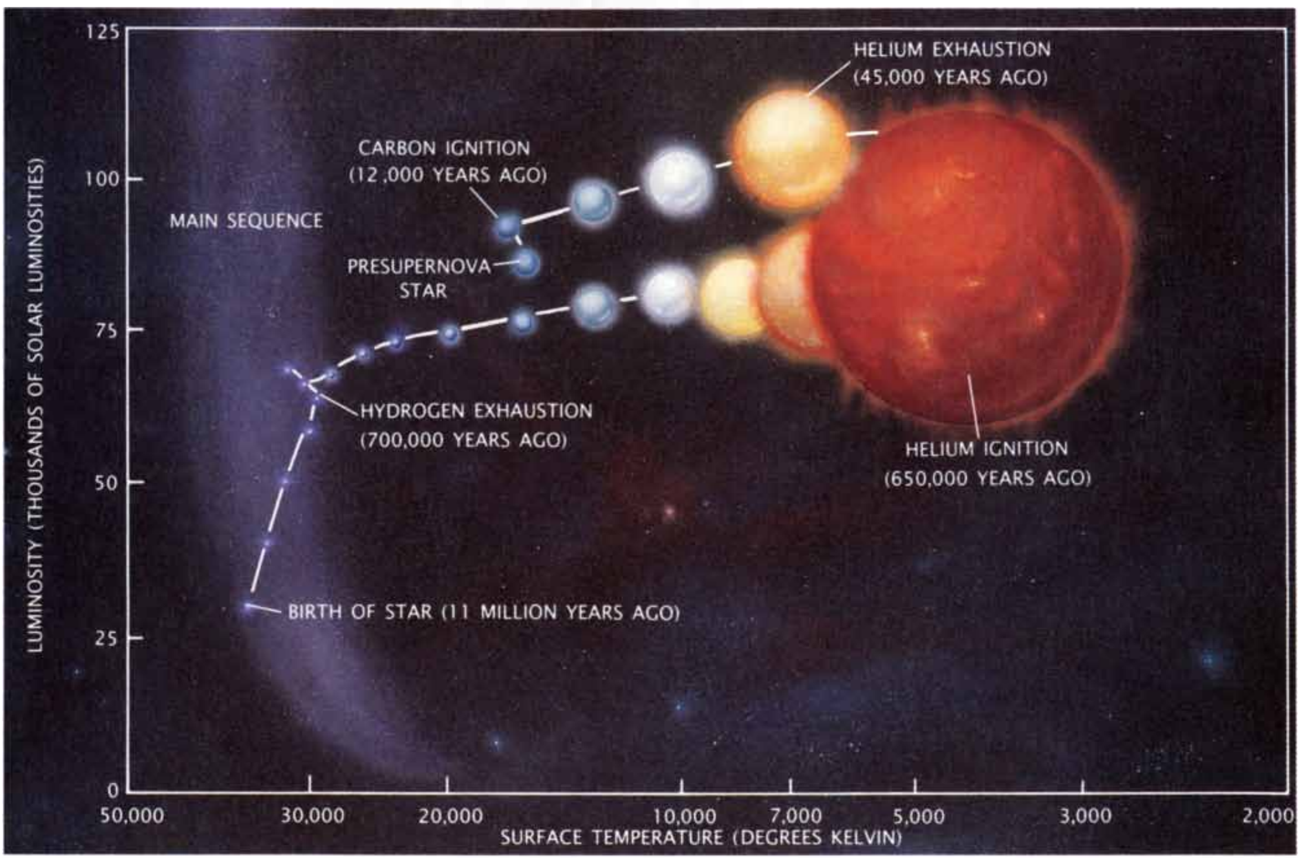
By heating and expanding the star and triggering a new flurry of nuclear reactions in its layered interior, the revived shock was responsible for the supernova's optical display. The effect was delayed by about two hours: the shock traveled at perhaps a fiftieth of the speed of light and had to traverse the entire star before any light leaked out. The neutrinos from the collapsing core easily outraced the shock. Passing through the rest of the star very close to the speed of light, they were the first signal to leave the supernova.

Some 160,000 years later, still hours ahead of the light front, the neutrinos swept over the earth—and were detected. Investigators searching for rare subatomic events such as the decay of the proton have built detectors deep in mines and under mountains, where they are shielded from interference by cosmic rays. Typically they consist of a swimming-pool-size tank of water flanked by arrays of photodetectors, poised to sense the faint flashes of light that would signal the decay of any one of the perhaps 10^{32} protons in the water. To date no proton has been seen to decay, but the detectors are also sensitive to another rare, energetic event, the capture of a neutrino by a proton.

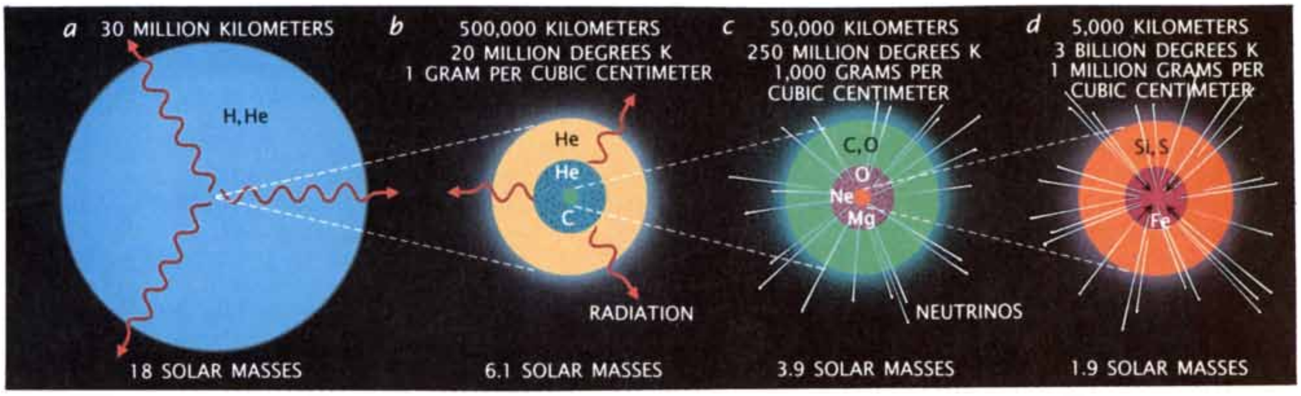
On February 23 at 7:36 A.M. Universal Time, the Kamiokande II detector, in the Kamioka lead mine in Japan, and the IMB detector (named for the collaborating institutions, the University of California at Irvine, the University of Michigan at Ann Arbor and the Brookhaven National Laboratory) in the Morton Thiokol salt mine near Cleveland, Ohio, simultaneously recorded a series of events that were later interpreted as neutrino captures. A detector of a different design, at Baksan in the Soviet Union, registered anomalous events at the same time. Approaching from out of the southern sky, the wave of neutrinos from the supernova had swept through the earth (the earth is far more transparent to these weakly interacting particles than a thin sheet of the clearest glass is to light). Emerging in the Northern Hemisphere, it had left the faintest signature of its passage in the detectors.

The theoretical significance of the neutrino detection was considerable. The Kamiokande and IMB detectors are most sensitive to a small component of the burst: electron antineutrinos. The same proportion of the burst energy is believed to have come from each of the other five neutrino flavors—electron neutrinos and mu- and tau-neutrinos and their two antiparticles. By extrapolating from the number and energy of the neutrinos that were detected, workers have calculated the total neutrino energy released by SN 1987A: 3×10^{53} ergs. It is just equal to the theoretical binding energy of a neutron star of 1.4 solar masses—the gravitational energy that should be released in its formation.

Thus, the fleeting detection of the neutrino burst shows that, as theory had predicted, a neutron star is formed in a type II supernova. More specifically, it is a sign that computer models of the formation and collapse of massive stars are on the right track: they had accurately predicted the mass of the imploding core. The average energy of the detected neutrinos confirms theoretical predictions for the temperature of a collapsing protoneutron star. Furthermore, the burst lasted several seconds: the neutrinos actually did have to diffuse out of the



HISTORY of the progenitor star began some 11 million years ago on the "main sequence"—the region hydrogen-burning stars (most stars in the sky) occupy on graphs of luminosity versus surface temperature. After about 10 million years the hydrogen in the star's core was burned to helium, and the core contracted and became hotter. In response the star's envelope expanded and cooled, and the star moved to the right, off the main sequence. As the core got hot and dense enough to burn helium, the star bloated into a red supergiant, with a cool envelope several times the size of the earth's orbit. After helium was exhausted, the envelope contracted and heated up again, and the star became a blue supergiant. In that form it burned successively heavier elements, ultimately making the iron core whose collapse triggered the supernova. The scenario is based on the authors' calculations for an 18-solar-mass star with a starting composition typical of the Large Magellanic Cloud.



COSMIC ONION—the structure of the presupernova star in its final moments—is made up of concentric shells of successively heavier elements undergoing nuclear fusion. The radius of each shell, the temperature and density at its surface and the mass the shell includes are given; stippling indicates regions undergoing convection. When the center of the star's vast, tenuous envelope of hydrogen and helium (*a*) is magnified by a factor of 30 (*b*), a core of helium four times the diameter of Jupiter is revealed. Enlargement by another factor of 10 (*c*) exposes the ash from helium-burning: a core of carbon and oxygen. The carbon is burning to make neon and magnesium, which, together with oxygen, are being turned into silicon and sulfur (*d*). A final step of fusion has turned silicon and sulfur into 1.4 solar masses of iron at the very center of the star. Iron is fusion's end point; lacking any way to maintain adequate heat and pressure, the Mars-size iron core is starting to collapse. At its center the density is 10 billion grams per cubic centimeter and the temperature is 10 billion degrees Kelvin.

dense matter of the collapsed core.

Of even broader significance was the fact that the neutrinos arrived as a close-packed bunch a few hours ahead of the light burst after a journey of 160,000 years. The universe is widely believed to contain far more mass than can be seen, and neutrinos have been proposed as the carrier of this "missing mass." The fact that the neutrinos traveled so close to the speed of light sets strict limits on their mass: neutrinos with significant mass traveling at such speeds would have been far more energetic than the detected particles. Furthermore, neutrinos of quite different energies arrived within seconds of one another; in contrast, the arrival times of particles with significant mass would have been spread out in order of decreasing energy.

Independent analyses of the timing by John Bahcall of the Institute for Advanced Study in Princeton, N.J., Adam Burrows of the University of Arizona and Tom Loredo and Don Lamb of the University of Chicago have yielded a firm upper limit on the electron antineutrino's mass: about 20 electron volts (.00004 times the mass of the electron). If the masses of mu- and tau-neutrinos could be limited to similar values, neutrinos could confidently be dismissed as a missing-mass candidate.

The neutrino burst bore tidings of the core collapse, but it had very little to say about how the shock generated by the collapse got out of the core. The revival of the shock by neutrino energy remains in the realm of theory. Nevertheless, it is clear that a strong shock did propagate through Sk −69° 202 on February 23, 1987 (minus 160,000 years). To state the obvious, the star did explode.

Two hours after the neutrinos had been registered in the Kamiokande and IMB detectors (nobody knew it at the time, of course), Albert Jones, an amateur astronomer in New Zealand, happened to be observing the exact spot at which the supernova would appear. He did not see anything unusual. An hour later, in Australia, McNaught exposed the two plates that, when they were developed after Shelton's announcement of the discovery, showed the earliest recorded light from the supernova. Sometime between the two observations, perhaps even as Jones observed the spot, the shock erupted through the surface of the star, triggering a hard (short-wavelength) ultraviolet burst that quickly gave way to visible light.

The fact that it took only about two hours after the core collapse for the shock to arrive at the surface and ignite the optical display helped to dispel initial doubts about whether the blue star Sk −69° 202 really was the star that exploded. The quick arrival of the light ruled out a red supergiant as the progenitor: it takes even a high-velocity shock the better part of a day to go through a red supergiant.

Further evidence about the size of the progenitor star came from the ultraviolet flash, even though only its aftermath was seen. In addition to being invisible, ultraviolet light is absorbed by the earth's atmosphere. The telescope on board the *International Ultraviolet Explorer* satellite could have detected this earliest light but was not aimed in the right direction at the time. Within 14 hours, however, the observing team, headed by Robert P. Kirshner of the Harvard-Smithsonian Center for Astrophysics and George Sonneborn of the National Aeronautics and Space Administration's Goddard Space Flight Center, had reoriented the satellite. By that time the initial burst was fading, but the supernova was still clearly visible at ultraviolet wavelengths.

Moreover, the workers got an indirect look at the ultraviolet flash months later, when the IUE detected emissions from a shell of gas surrounding the supernova at a distance of about a light-year. The gas, presumably material ejected from the presupernova star in a stellar wind during its red-supergiant stage 40,000 years before, was flash-ionized when the intense ultraviolet burst reached it. Based on this secondary radiation, Claus Fransson of the University of Stockholm concluded that the first light from the supernova came from material at a temperature of about half a million degrees K. (Perhaps 10 years from now, according to a model developed by Roger A. Chevalier of the University of Virginia, the shell will radiate again, this time in the radio and X-ray bands, when the supernova ejecta finally collide with it.)

Such high temperatures, and the

very hard ultraviolet radiation they produce, are expected when a powerful shock wave breaks through the surface of a relatively small progenitor star. With less surface area in which to deposit its energy, the shock generates a correspondingly higher temperature, and it also accelerates the material to higher velocities. Doppler-shifted lines in the early ultraviolet and optical spectra indicated that the material had been ejected from the star at roughly one tenth the speed of light.

This expansion cooled the outermost layers of the young supernova, and the dominant emissions quickly shifted from the ultraviolet to the cooler, visible wavelengths recorded in the earliest photographs. The bolometric luminosity (the combined radiation at all wavelengths from the infrared through the ultraviolet) declined steeply during these first hours, but because the visible portion of the emissions was strengthening, the supernova was brightening into an impressive display in the night sky.

During the first day or so, little radiation could escape from deep inside the supernova: free electrons in the ionized gas of its envelope scattered light from deeper layers. When the outermost material had cooled to about 5,500 degrees K, however, the hydrogen nuclei recombined with the free electrons. As the supernova continued to expand and cool, a surface defined by the hydrogen recombination temperature moved into the envelope. At this surface energy previously deposited by the shock was released—mostly at visible wavelengths—and streamed freely into space. For weeks to come, as Arnett and Sydney W. Falk of the University of Texas at Austin had predicted some 15 years ago, radiation at the hydrogen recombination temperature dominated the supernova's emissions.

At the same time another effect of the small progenitor star became apparent. As an optical display the supernova was at first unexpectedly faint—about a tenth as bright as other type II supernovas at a similar stage. To cool to the hydrogen recombination temperature, any supernova has to expand. The shock had deposited about the same amount of energy in the relatively small envelope of this progenitor star as it would have left in a red supergiant's extended envelope, heating the small envelope to a correspondingly higher temperature. As a result SN 1987A had to expand by a much larger factor before it could release its light, and the process consumed energy that would otherwise have come out as radiation.

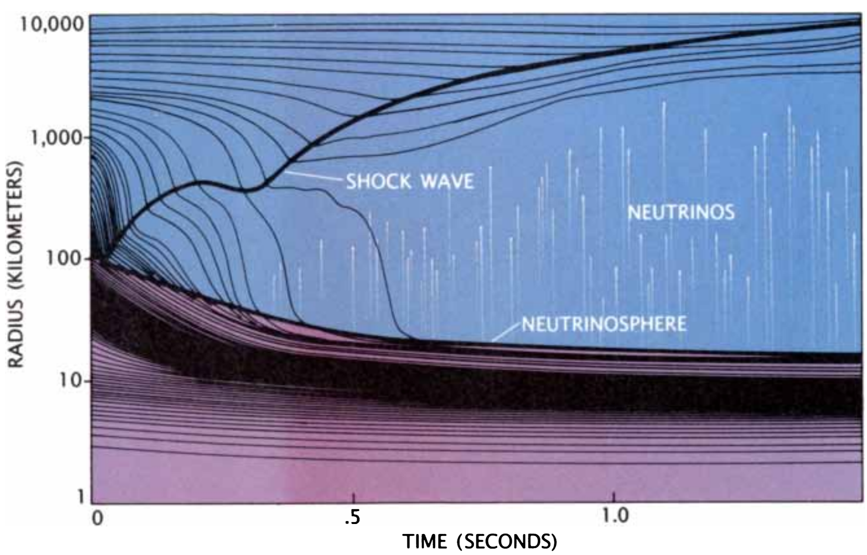
After about a month, it is calculated, all the energy deposited by the shock had either escaped as radiation or gone into accelerating the ejecta. Yet the supernova was still brightening at visible wavelengths. By this time, in April, another source of energy was providing most of the light: the decay of radioactive isotopes produced in the explosion. Most theorists had expected such materials to be made in a type II supernova, but they watched eagerly to see how much had been generated in SN 1987A and what role the isotopes would play.

The shock wave's passage through deep layers of the progenitor star during the first minutes of the event had triggered new nuclear reactions. In particular, part of the silicon shell was turned into iron-group elements, chiefly the radioactive isotope nickel-56. A month later the highly unstable nickel-56 had already decayed (its half-life is 6.1 days), heating and expanding the deep layers of the supernova. But its decay product, cobalt-56, is also radioactive, and because it has a half-life of 77.1 days, it was still abundant. It decays into an excited iron-56 nucleus, which releases gamma rays at specific energies as it relaxes to the ground state. These gamma rays now powered the display.

At first the gamma rays themselves did not escape: because of their high energy they scattered repeatedly from electrons in the expanding gas, turning into X rays of progressively lower energy. At a sufficiently low energy the X rays were absorbed, heating the material and so contributing to the optical display. As the supernova continued to thin, increasing amounts of the decay energy escaped in this way. On May 20, 80 days after the explosion, the brightness peaked.

By early July the light curve was declining at precisely the rate expected on the basis of cobalt-56's half-life. From the brightness of the supernova on a given day and the time since the explosion, it was straightforward to calculate how much nickel-56 had been formed in the first place. The answer, .08 solar masses, is within a factor of two of what we and others had predicted for type II supernovas.

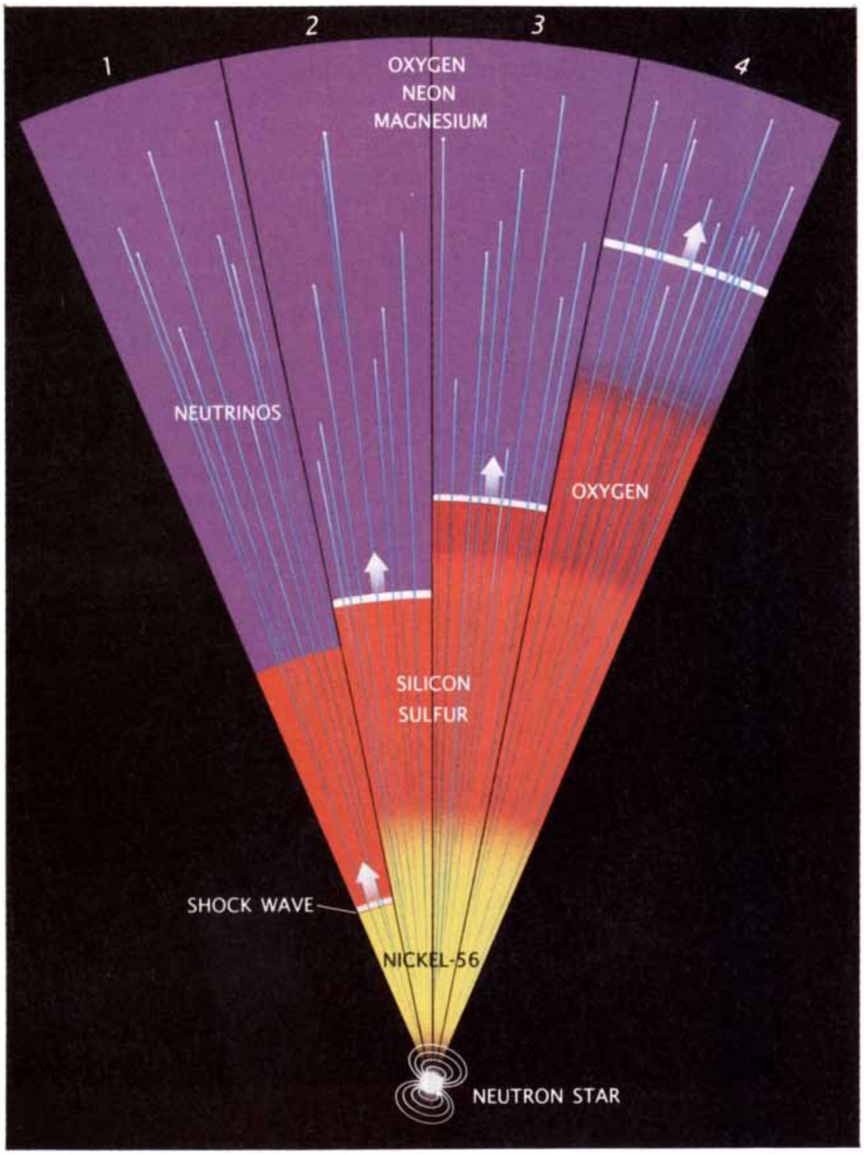
For weeks after peak brightness the radioactive material still could not be seen directly. By August, however, the expanding debris had thinned enough to allow some radiation from the decay to escape with little or no scattering. First the Japanese satellite *Ginga* and, shortly thereafter, instruments on the Soviet space station *Mir* detected X rays at the energies that Philip A. Pinto of the University of California at Santa Cruz, Rashid A. Sunyaev and S. A. Grebenev of the Soviet Space Re-



NEUTRINOS REVIVE THE SHOCK WAVE generated by core collapse in a simulation calculated for a star about the mass of supernova 1987A's precursor. Each line on the graph traces the radial position of a shell of constant mass. As the time scale begins, the shock has lost energy and stalled, a few hundredths of a second after its birth, within the infalling material of the outer core. The graph shows how the collapsed core (*purple*)—a protoneutron star—contracts further and emits a powerful flux of neutrinos, which escape from its surface (the "neutrinosphere") after diffusing through the nuclear matter. A trace of energy deposited by neutrinos heats and accelerates material behind the shock. The revived shock is sufficient to destroy the star. The calculations were done by Ron Mayle and James R. Wilson of Livermore.

search Institute, and others had predicted would be seen when the gamma rays from cobalt-56 decay were scattered. Once the X rays had been seen, the gamma rays themselves could not be far behind, and in December their discovery was announced based on data from the *Solar Maximum Mission* satellite. Confirmation came quickly from balloon-borne detectors flown in Australia and in Antarctica.

Donald D. Clayton of Rice University and his co-workers had predicted some 20 years ago that a supernova would produce gamma rays of the observed energies, but the timing of the observations was a surprise. Theorists had expected that in a type II supernova the layers of the exploded star would expand in radial symmetry, in which case the X rays would have been obscured until perhaps 100 days after they were actually observed. Their early appearance meant, instead, that the core had been mixed: material from the inner layers had been blown into the overlying layer of helium or even into the hydrogen envelope. Indeed, Doppler broadening of the gamma-ray lines showed that some cobalt was moving as fast as 3,000 kilometers per second—fast enough to have overtaken the slower-moving material at the base of the hydrogen envelope.



EXPLOSIVE NUCLEOSYNTHESIS takes place as the shock wave rips out through the progenitor star's layered interior. Shock-heated to more than five billion degrees K, part of the silicon and sulfur fuses to form radioactive nickel-56 (*stage 1*); some of the oxygen at the bottom of the next shell burns to silicon and sulfur (*stage 2*), and neon and magnesium in the inner part of the shell burn to oxygen (*stage 3*). The shock propagates through the remaining material without triggering further transmutations (*stage 4*). The decay product of the nickel, cobalt-56, is also radioactive and yields much of the energy for the supernova's light. Neutrinos from the hot, contracting neutron star at the center of the supernova outrace the shock wave.

At about the time the cobalt appeared, emissions from deeper in the supernova revealed other heavy elements. Gamma rays and X rays from the core of the supernova were still being scattered, and visible and ultraviolet emissions were blocked by a thicket of atomic absorption lines. The infrared, it turned out, offered the earliest look at the heavy elements the supernova was dispersing into space.

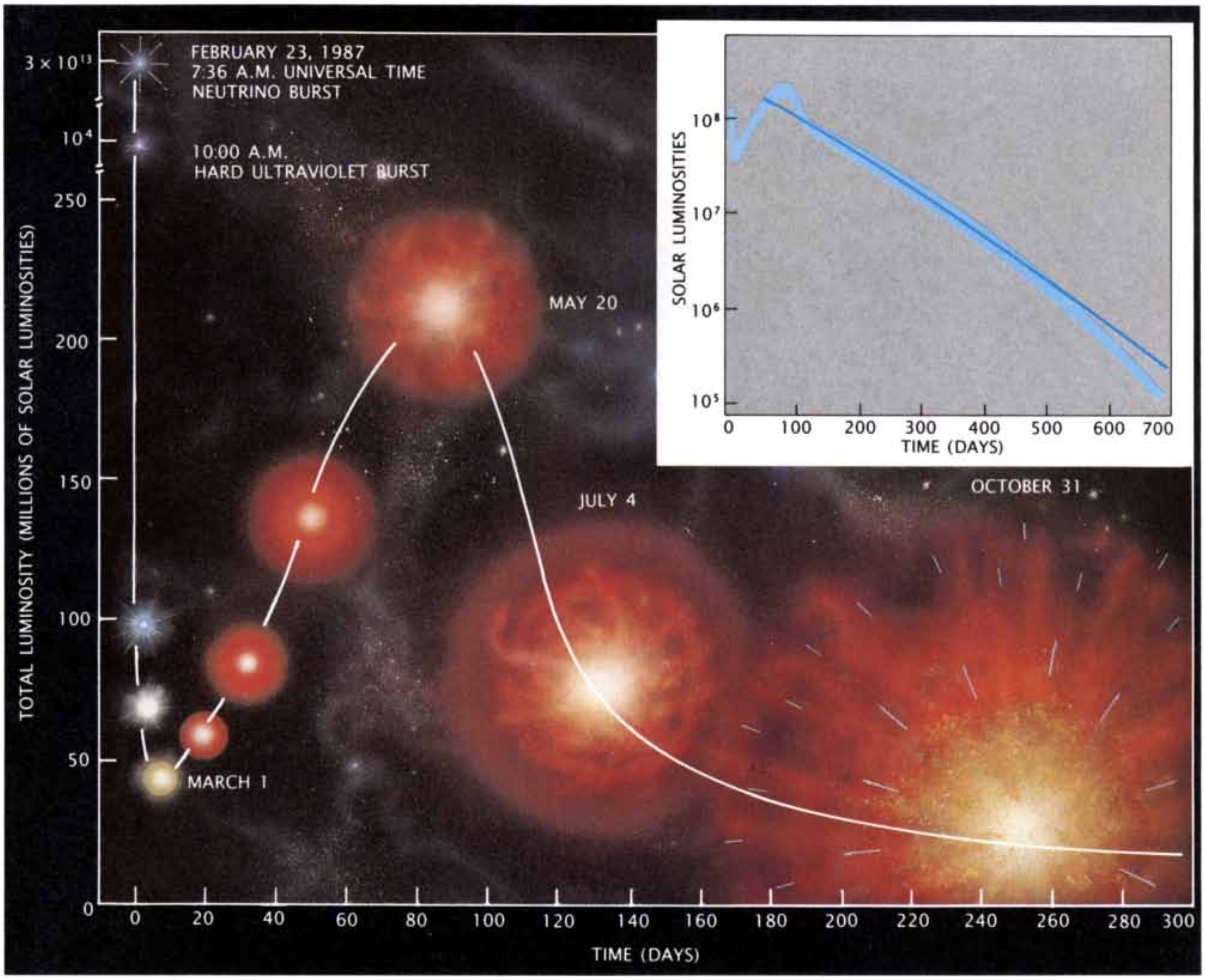
Most infrared radiation is absorbed by the earth's atmosphere, but the wavelengths that do reach the ground were studied, beginning soon after the supernova exploded, by the Anglo-Australian Telescope at Coonabarabran and the Mount Stromlo and Siding Spring Observatories in Woden (both in Australia) and by the Cerro Tololo Inter-American Observatory in Chile. NASA's Kuiper Airborne Infrared Telescope, flown at 39,000 feet on a jet transport, gained more complete coverage on flights starting in the fall of 1987. Beginning around November, spectra from the Kuiper and from Australia together revealed an entire zoo of elements in the supernova core—not just iron, nickel and cobalt but also argon, carbon, oxygen, neon, sodium, magnesium, silicon, sulfur, chlorine, potassium, calcium and possibly aluminum. Their intense infrared lines signaled larger quantities than could have been present in the star at its birth. The elements—the components, perhaps, of some future solar system—were made in the core of the star or in the explosion itself.

In early 1989, two years after the explosion, the supernova's luminosity was declining steadily, in keeping with the exponential decay of radioactive cobalt-56 (save that some of the X rays and gamma rays could now escape directly and hence did not contribute to the light curve). The lack of evidence for any energy source other than radioactive decay was starting to puzzle some theorists. The neutrino burst had announced the birth of a neutron star. Yet a neutron star usually emits a great deal of radiation, either by heating any material falling into it or by acting as a pulsar: a spinning neutron star with a strong magnetic field that generates a rotating beacon of radiation.

Where was the neutron star in SN 1987A? Had it formed initially but then vanished by turning into a black hole? The neutrino burst would have been cut short if a black hole had formed during the first few seconds of the event, and in any case the mass of the collapsing iron core alone fell short of the threshold—about two solar masses—for forming a black hole. If enough additional mass had later fallen onto the neutron star to drive it over the limit, all the radioactive nickel would have been lost and the supernova would have been much fainter. As the supernova neared its second anniversary, most astronomers were still betting on a neutron star, although the exponential decline of the light curve ruled out a very bright pulsar such as the one in the Crab Nebula, the remnant of a brilliant supernova in 1054.

During the night of January 18, 1989, Universal Time, the supernova answered one puzzle with several more. At Cerro Tololo a group headed by Carl Pennypacker of the Lawrence Berkeley Laboratory and John Middleditch of Los Alamos detected optical pulsations from the supernova. The pulsations, which amounted to about .1 percent of the total light, came nearly 2,000 times a second, suggesting a rotation rate three times faster than the fastest pulsar ever seen. Spinning that fast, only the densest, most massive neutron star allowed by theory could avoid flying apart.

What is more, the signal of the pulsar showed a regular variation in frequency, as if an orbiting companion several times as massive as Jupiter were tugging the pulsar back and forth every seven hours, Doppler-shifting its signal. Because the radius of the companion's calculated orbit, about a million kilometers, would have placed it inside the presupernova star, the companion could only have been created after the explosion. Spec-



EXPLOSION of the supernova began with an enormously powerful burst of neutrinos, marking the birth of the neutron star, followed some two hours later by a flash of hard ultraviolet light as the shock wave broke through the surface of the star, heating it to half a million degrees K. Over the next few days the surface of the supernova expanded and cooled to red heat. Energy deposited by the shock deep inside the envelope began leaking out, joined several weeks later by heat from the decay of radioactive cobalt. The supernova brightened slowly until May 20, by which time the shock energy had been spent and the display was powered entirely by radioactivity. The subsequent decline in brightness, plotted on a logarithmic scale (*upper right*), matched the decline calculated for the decay energy of .08 solar masses of cobalt-56 (*dark curve*). Months after the explosion, as the supernova thinned into a clumpy nebula many times the size of the solar system, X rays and gamma rays (*blue arrows*) from the decay of the cobalt began to escape directly. Even today the supernova is an unresolved point of light in telescopes; the paintings are based on each stage's observed color and spectrum and on theoretical inferences.

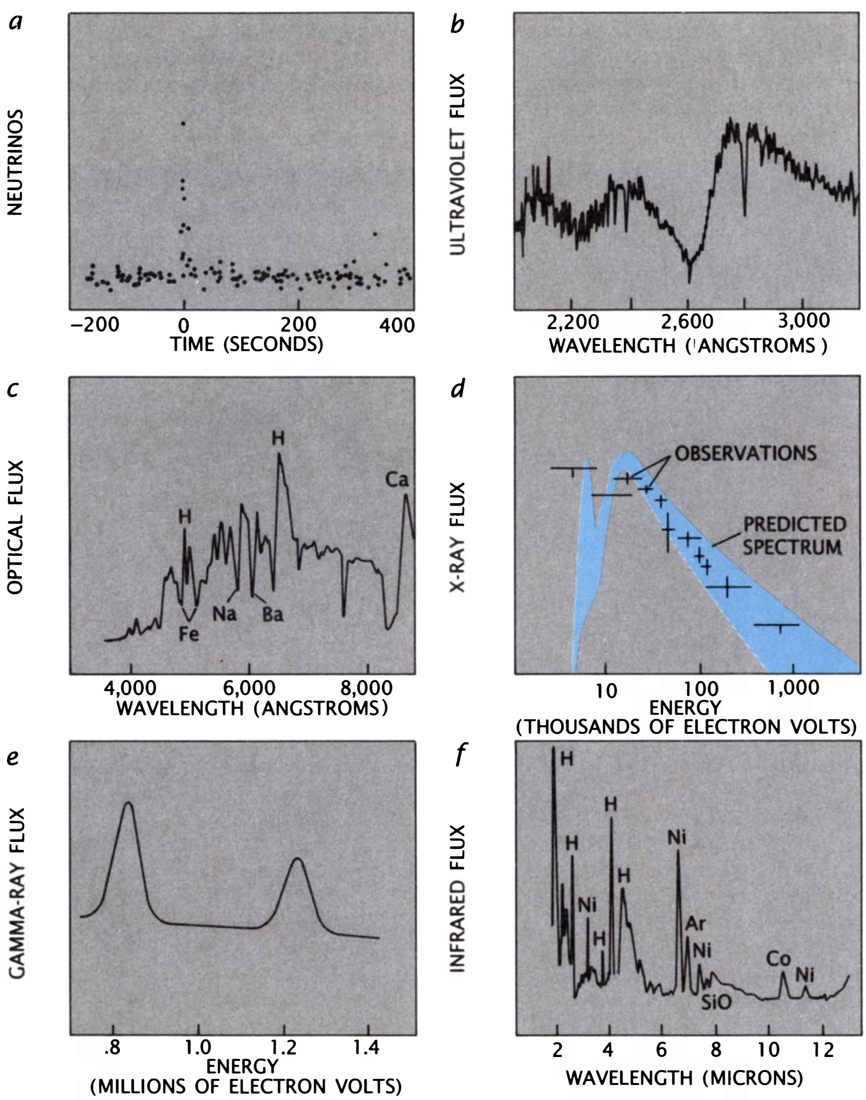
ulation is rife: if the companion is real, could it be a piece of neutron star that was somehow ejected, some other fragment that fell back and was captured, or something even more exotic?

What is really needed is another look at the pulsar. Yet several months of observations of equal and greater sensitivity have failed to recover it. Again, one can speculate. Clouds deep in the supernova may be obscuring the pulsar, for example, or it may have been extinguished: matter falling onto the neutron star may have short-circuited the electric field (generated by the rotating magnetic field) that powers the beam. No one knows.

The fate of the neutron star joins other mysteries that have accompanied SN 1987A. We have emphasized the success of theory and the beautiful complementarity among the observations. Yet there had been anomalies even before the putative pulsar. Four hours *before* the neutrino detection at Kamiokande and IMB, for example, a detector in a tunnel under Mount Blanc had registered a separate neutrino burst. Gravity-wave detectors (sensitive to massive releases of gravitational energy) in Rome and in Maryland are said to have recorded signals coincident with those early neutrinos. What could account for a stupendous burst of energy four hours before the core collapse? Again, no one knows. Several months after the explosion came another mystery: a second light source, roughly one tenth as bright as the supernova and resolvable from the main explosion only by an indirect technique known as speckle interferometry. The mysterious second source had disappeared by June, 1987, and was not seen again.

Doubts about such observations, and controversy about their interpretation, bring home an important point about the supernova. In much of science a result is accepted only if it is reproducible. Yet in the case of supernova 1987A we deal with an event that may not be repeated nearby for centuries. When our ability to interpret the observations breaks down, the best we can do is to record and archive the findings carefully, so that future scientists, with greater insight, may come to understand them.

Even so, the last two and a half years have yielded breathtaking advances in the understanding of type II supernovas. For us and hundreds of others, theorists and observers at all wavelengths collaborating to document and explain one of the heavens' grandest events, it has been a time of matchless exhilaration, scientific cooperation and intellectual reward—the event of a lifetime.



EMISSIONS from SN 1987A began with a brief burst of neutrinos, shown in a record from the Kamiokande detector in Japan (*a*). Hours after the shock wave burst from the star, the *International Ultraviolet Explorer* satellite recorded an ultraviolet spectrum testifying to the very high temperature of the shock-heated surface (*b*). A spectrum at visible wavelengths, made 50 days after the explosion, shows strong spectral lines of hydrogen, characteristic of the expanding, cooling envelope (*c*). After about six months instruments on the Japanese satellite *Ginga* and the Soviet space station *Mir* detected X rays from the decay of radioactive cobalt (*d*); the detection of gamma rays from the same decay by the *Solar Maximum Mission* satellite was reported a few months later (*e*). Infrared emission lines captured by the National Aeronautics and Space Administration's Kuiper Airborne Infrared Telescope reveal a variety of newly made elements deep in the expanding ejecta (*f*). The ultraviolet spectrum was provided by Robert Kirshner of the Harvard-Smithsonian Center for Astrophysics, the gamma-ray data by Mark D. Leising of the Naval Research Laboratory and the infrared spectrum by Fred C. Witteborn of the NASA Ames Research Center.

FURTHER READING

THE PHYSICS OF SUPERNOVA EXPLOSIONS. S. E. Woosley and Thomas A. Weaver in *Annual Review of Astronomy and Astrophysics*, Vol. 24, pages 205-253; 1986.

SUPERNOVA 1987A! S. E. Woosley and M. M. Phillips in *Science*, Vol. 240, No. 4853, pages 750-759; May 6, 1988.

SUPERNOVA 1987A IN THE LARGE MAGELLANIC CLOUD. M. Kafatos and A. Michalitsianos. Cambridge University Press, 1988.

THE SUPERNOVA STORY. Laurence A. Marschall. Plenum Press, 1988.

SUPERNOVA 1987A. W. David Arnett, John N. Bahcall, Robert P. Kirshner and Stanford E. Woosley in *Annual Review of Astronomy and Astrophysics*, Vol. 27, pages 629-700; 1989.

Introducing...

CARE *of the* SURGICAL PATIENT

from SCIENTIFIC AMERICAN *Medicine*

Because the quality of your care depends on the quality of your information.

Treating pre and post operative patients poses a unique set of challenges. Yet in one way it's no different than any other practice issue.

Doing it well takes the right information.

That's why SCIENTIFIC AMERICAN *Medicine* is pleased to announce the publication of *CARE of the SURGICAL PATIENT.*

The definitive resource on pre and post-operative care.

CARE of the SURGICAL PATIENT gives you ready access to the most authoritative and current information on pre and post-operative standards available anywhere.

Written and designed by prominent surgeons under the supervision of the American College of Surgeons' Committee on Pre and Postoperative Care, *CARE of the SURGICAL PATIENT* provides two volumes – over 1,500 pages – of practical information on both critical and elective care.

And, *CARE of the SURGICAL PATIENT* is updated twice a year, with each surgeon-author reviewing his own specialty. Updates include new information on significant topics, such as current developments on AIDS.

In short, *CARE of the SURGICAL PATIENT* presents the standards for pre and postoperative treatment. You simply won't find a more important resource. Or one that organizes its information in such an intelligent way.

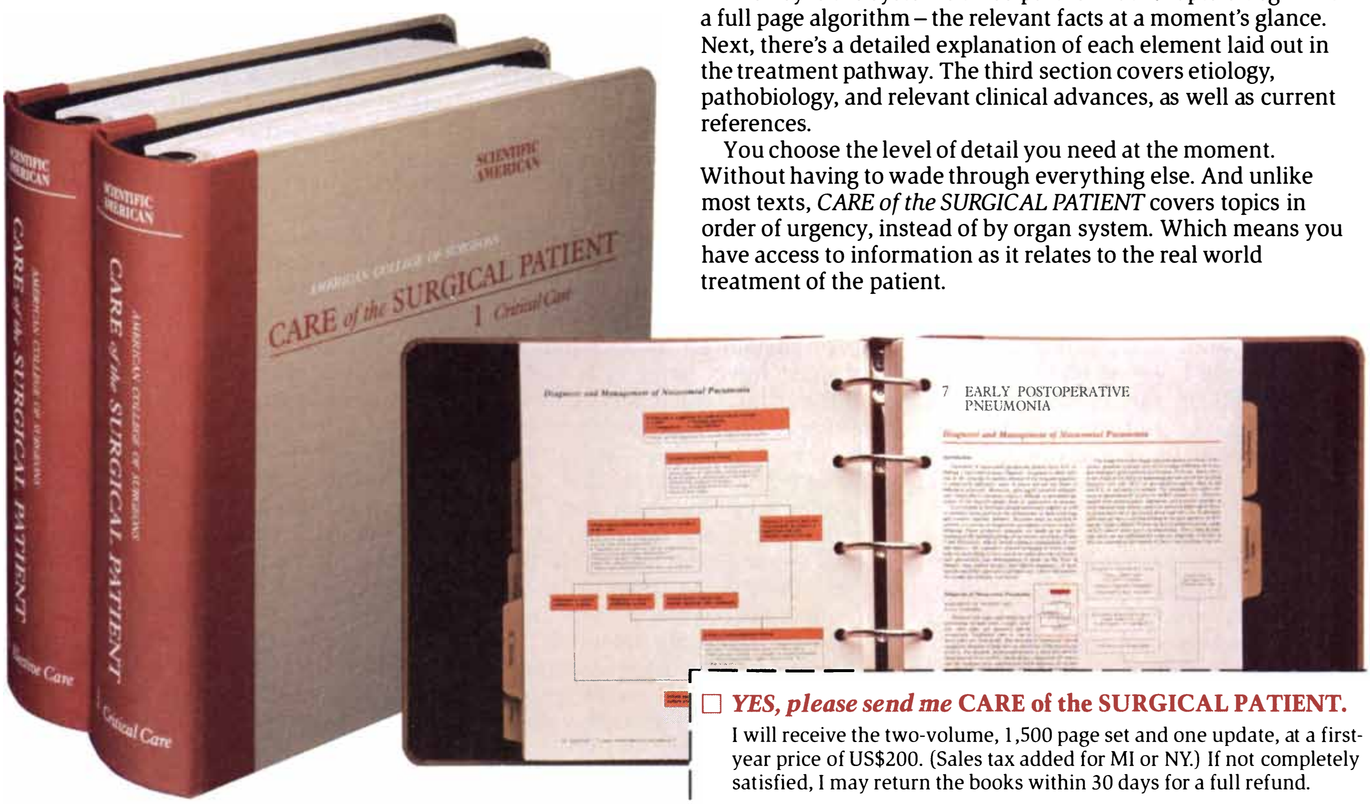
A unique system for rapid information retrieval.

CARE of the SURGICAL PATIENT is designed to get you the information you need, the way you need it.

Quickly. And intelligently.

The key is the system's three-part format. Chapters begin with a full page algorithm – the relevant facts at a moment's glance. Next, there's a detailed explanation of each element laid out in the treatment pathway. The third section covers etiology, pathobiology, and relevant clinical advances, as well as current references.

You choose the level of detail you need at the moment. Without having to wade through everything else. And unlike most texts, *CARE of the SURGICAL PATIENT* covers topics in order of urgency, instead of by organ system. Which means you have access to information as it relates to the real world treatment of the patient.



Try CARE of the SURGICAL PATIENT *Free for 30 days.*

You'll find it the most valuable resource on pre and post-operative care that's ever been published. And if you're not satisfied, just return it. No risk. No obligation.

CARE of the SURGICAL PATIENT, from SCIENTIFIC AMERICAN *Medicine*. No other resource helps you keep up better.

And the better you keep up, the better your care.

☐ ***YES, please send me*** **CARE of the SURGICAL PATIENT.**

I will receive the two-volume, 1,500 page set and one update, at a first-year price of US$200. (Sales tax added for MI or NY.) If not completely satisfied, I may return the books within 30 days for a full refund.

☐ Check enclosed ☐ MasterCard ☐ VISA ☐ Bill me

Acct.# ______________________ Exp. Date ________

Name ______________________

Address ______________________

____

City ______________ State ________ Zip ________

Specialty ______________________

Or call toll free 1-800 345 8112.

SCIENTIFIC AMERICAN MEDICINE 415 Madison Avenue, New York, NY 10017 7506

Viral Alteration of Cell Function

Certain viruses interfere subtly with a cell's ability to produce specific hormones and neurotransmitters. Persistent infections by such viruses may underlie a multitude of glandular and organic disorders

by Michael B. A. Oldstone

"It has been well said that a virus is a piece of bad news wrapped up in protein," wrote Peter and Jane Medawar a few years back. The description is apt: when a virus infects a cell, the viral genes—the "bad news"—can, and often do, disturb the cell's normal activities. The injured cells may then die or malfunction, which gives rise to disease in the organism. Indeed, it was through the visible devastation viruses wrought in plants and animals that science first became aware of their existence.

Today a pathologist or virologist who suspects a viral infection looks with a microscope for infected cells leaking to death, their membranes riddled with escaping virus particles, as lymphocytes and other immune cells close in like riot police to contain the infection. Such immune cells themselves can damage tissues. If one finds these indicators, one then searches the patient's blood, urine and tissues for viral antibodies, genes and proteins in the hope of identifying the virus itself. In the absence of these hallmarks, the problem is usually assumed not to be viral in origin.

Within the past decade, however, a number of investigators, including my group at the Scripps Clinic, began to turn up evidence suggesting that the situation may not always be so clear-cut. We have found viruses that reside in cells and yet do not produce the classic hallmarks of viral infection: the viruses do not kill the cells, and at the same time they do not elicit an effective immune response. These tactics enable the viruses to establish a long-term presence within cells, where they can have a subtle and persistent effect—often by altering a specialized function of the cell, such as the production or secretion of a hormone. Such "luxury" functions are not essential to the cell's survival but may be vital to the health of the organism. There is now increasing evidence that this insidious mode of viral activity may underlie many human illnesses, such as certain kinds of growth retardation, diabetes, neuropsychiatric disorders, autoimmune disease and heart disease, that have not been suspected to have an infectious cause.

That this should be so is not surprising if one pauses to reflect on the singular nature of viruses. Viruses are genetic parasites, unable to sustain an independent life; they must infect living cells and exploit the cellular machinery in order to replicate. Within this limitation, however, viruses have evolved a diversity of survival strategies—and equally diverse ways in which they cause disease. There are, first of all, the agents—such as the poliomyelitis virus and the common cold virus—that can cause acute infection and rapid cell death. There are latent viruses, such as the notorious herpes simplex, which can lie dormant and undetected in cells for long periods before flaring up, painfully. Another type of chronic infection is caused by "slow viruses," which build up gradually over a long period before giving rise to illness; slow viruses have been implicated in two rare human diseases, kuru and Creutzfeldt-Jakob, both marked by progressive dementia.

Evidence of yet another mode of persistent viral activity came from investigations by my group and others of lymphocytic choriomeningitis virus (LCMV), a disease that is endemic in certain wild mouse populations. Studied intensively since the 1930's, LCMV has been a Rosetta stone for deciphering the mechanisms of viral infection and immune response. It was known that the virus could persistently infect cells without killing them. Such studies had been conducted on fibroblasts, however—a type of cell that does not have particularly interesting specialized functions. My colleagues and I wondered what effect LCMV would have on the specialized functions of differentiated cells.

With this in mind, we decided to infect neuroblastoma cells, which produce enzymes that make and break down the neurotransmitter acetylcholine. We noted that the infected cells developed abnormalities in the synthesis and degradation of acetylcholine. Yet the cells continued to grow normally and to produce normal levels of total DNA, RNA, protein and vital enzymes. Under the microscope the infected cells were indistinguishable from uninfected ones.

Howard Holtzer and his colleagues at the University of Pennsylvania found a corresponding situation in differentiating chick cells infected by a variant of the Rous sarcoma virus. The virus variant replicated only at certain temperatures, so that one could specify the effect of the virus on the cells by shifting the temperature. At temperatures unsuitable for viral replication the cells continued to produce the normal complement of their specialized products. When the temperature was shifted to permit viral replication, however, the cells ceased to manufacture their characteristic products. The morphology of the cells also changed, but they continued to carry on the activities they depend on to survive.

Subsequently a number of investigators found that other viruses also interfered with specialized functions of cells without disturbing their vital functions. Certain human and ani-

MICHAEL B. A. OLDSTONE is a member of the departments of immunology and neuropharmacology at Scripps Clinic and Research Foundation in La Jolla, Calif., where he heads a laboratory of viral immunobiology. Oldstone received a B.S. from the University of Alabama in 1954 and got his M.D. from the University of Maryland School of Medicine in 1961. He joined Scripps in 1969 and since 1972 has served as adjunct professor of pathology and neuroscience at the University of California, San Diego.

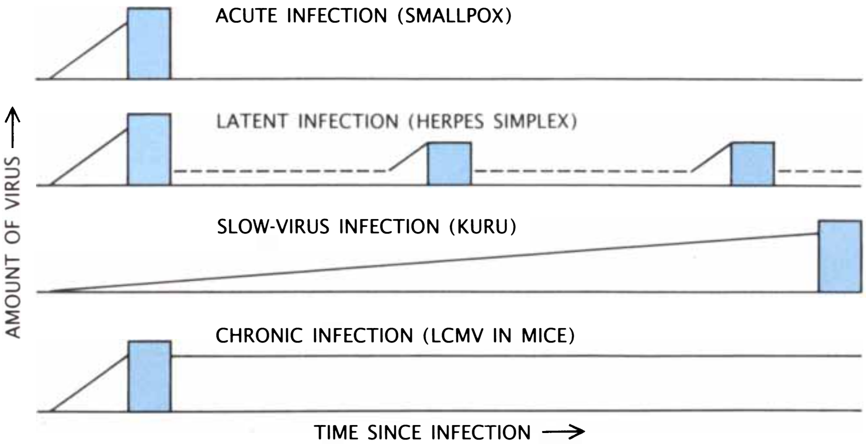
mal viruses were shown to infect and thereby alter the specialized functions of neurons (nerve cells), glia (a type of brain cell) and lymphocytes. It is interesting that the various viruses observed to infect lymphocytes were shown both to prevent the synthesis by *B* cells of immunoglobulin (antibody) molecules and to interfere with the ability of cytotoxic *T* cells to destroy infected cells.

These novel observations were made in the test tube. We wondered whether they would also pertain to whole animals. The first affirmative evidence came from our studies of a particular strain of mice (called C3H) that is chronically infected with LCMV. These mice had retarded growth and hypoglycemia (low blood glucose). Why should this be? Since both effects are regulated by growth hormone, we decided to examine the cells in the pituitary gland that make and secrete growth hormone. We found that the virus had invaded the anterior lobe of the pituitary and, what is more, was replicating preferentially in cells producing growth hormone. The result was that the animals produced 50 percent less growth hormone than is normal.

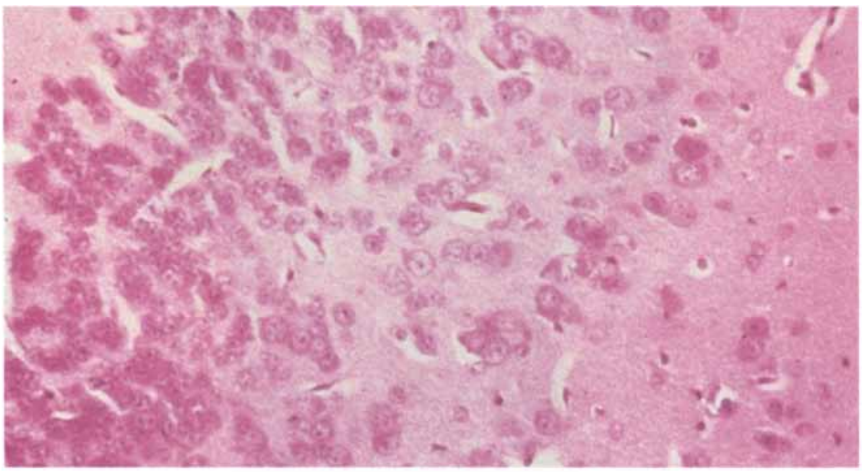
When we examined the infected cells under the microscope, we saw no evidence of cell injury or inflammation. Were the stunted growth and aberrant glucose metabolism caused by a malfunction of these cells or by some other, undiscovered defect? To find out, we transplanted healthy, uninfected growth-hormone-producing cells into the infected mice to see whether the transplanted cells would restore normal growth-hormone levels. The operation was a success: the growth rate and glucose metabolism of the mice returned to normal after the transplants. We had proved that infection by LCMV had interfered with the ability of the pituitary cells to synthesize growth hormone. Although the telltale footprints of viral infection—cell injury and inflammation—were absent, the virus had disturbed homeostasis and caused disease.

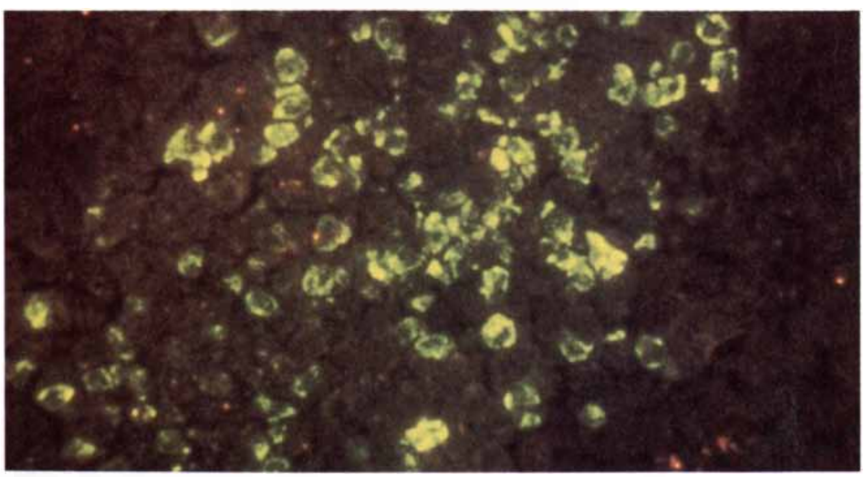
Having determined that LCMV infection leads to a growth-hormone deficiency, we wanted to learn the underlying biochemical cause of the deficiency. We first looked for a defect in the protein structure of the hormone produced by infected mice by comparing it with that of the growth hormone produced by uninfected mice. Since it is difficult to carry out a sequence-by-sequence comparison, we chopped up the protein into its component pep-

DIMINUTIVE SIZE of the midget Major White, shown held aloft by the giant James Tarver in this 1926 photograph, was caused by a deficiency of growth hormone. The disorder is genetic in some cases, but in others the cause is unknown. The author's discovery of a persistent virus that interferes with the production of growth hormone in mice suggests that it would be worth examining growth-hormone deficiency and other hormonal and neurotransmitter disorders for possible viral origins.



VIRAL INFECTION has several stages, which vary in duration and severity depending on the virus. After the initial infection there is an incubation period during which the amount of detectable infectious virus (*solid line*) increases until it brings on disease symptoms (*box*), including cell injury and inflammation. Latent viruses may cause an initial acute infection and then lapse into a quiescent phase during which the virus is not readily detectable (*broken line*); the virus can later flare up and cause a recurrence of disease. In contrast, persistent viruses can initiate chronic infections in which the virus is detectable but does not elicit a sufficient immune response to clear the virus. Recent findings suggest that such chronic infections can disturb cell function and give rise to disease in the absence of classic indicators.





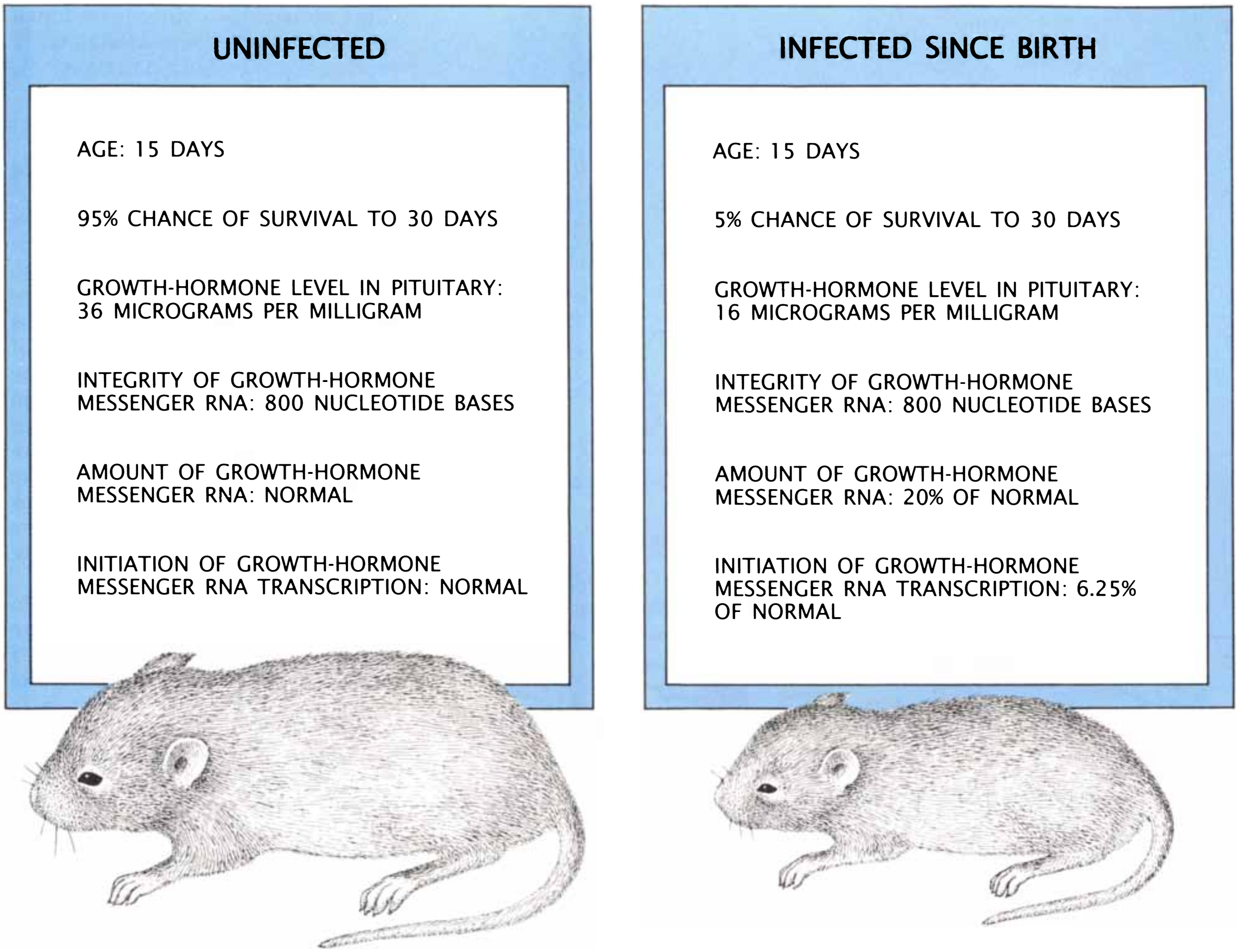
NEURONS from a mouse with a persistent lymphocytic choriomeningitis virus (LCMV) infection appear outwardly normal in an ordinary photomicrograph (*top*). The presence of the virus is revealed in a fluorescence photomicrograph (*bottom*); the fluorescent dye is delivered to infected cells by an antibody to an LCMV protein.

tides and compared them instead. The result showed that the growth hormone in both infected and uninfected mice had the same components. Because protein is made according to information encoded on a strand of messenger RNA, we next examined the growth-hormone messenger RNA and found it to be the same length in both infected and uninfected mice. It was therefore unlikely that the defect lay in the protein or RNA structure of the hormone itself.

We next wondered whether infected cells were simply unable to manufacture enough of the hormone. Such a shortage could be caused by a deficiency in the amount of messenger RNA. This supposition turned out to be correct when we analyzed the RNA in the pituitary cells: we found that the level of growth-hormone messenger RNA in infected mice was only one fifth of the level in age- and sex-matched uninfected controls. The information on messenger RNA is transcribed from the DNA of a gene, and so we theorized that the flaw might be in the transcription of growth-hormone DNA into messenger RNA. And indeed, such was the case: transcription of the gene was initiated only one sixteenth as much in infected mice as in the controls. This defect was unique to the growth-hormone gene; transcription proceeded normally for other genes, such as those for thyroid-stimulating hormone and actin (an essential structural protein), in both infected and uninfected mice.

To uncover further details about the precise mechanism by which LCMV interfered with transcription of the growth-hormone gene, we needed to know more about the virus itself. What viral genes were responsible for the interference? In searching for the answer, we were helped by the fact that there are two strains of LCMV, one (Armstrong) that causes disease and one (WE) that does not. LCMV has two RNA segments, a short and a long one. Yves Rivière, Rafi Ahmed and I took the two strains and swapped short and long segments among them and injected the reassorted viruses into mice. Only the viruses that contained the short segment of the Armstrong strain caused disease.

The next logical step would be to isolate the genes that are responsible for disease. We knew that the short strand is made up of 3,700 nucleotide bases and that one end encodes the viral glycoproteins (which form the outer coats of virus particles) and the other encodes the core proteins. To find out which part is the culprit, we

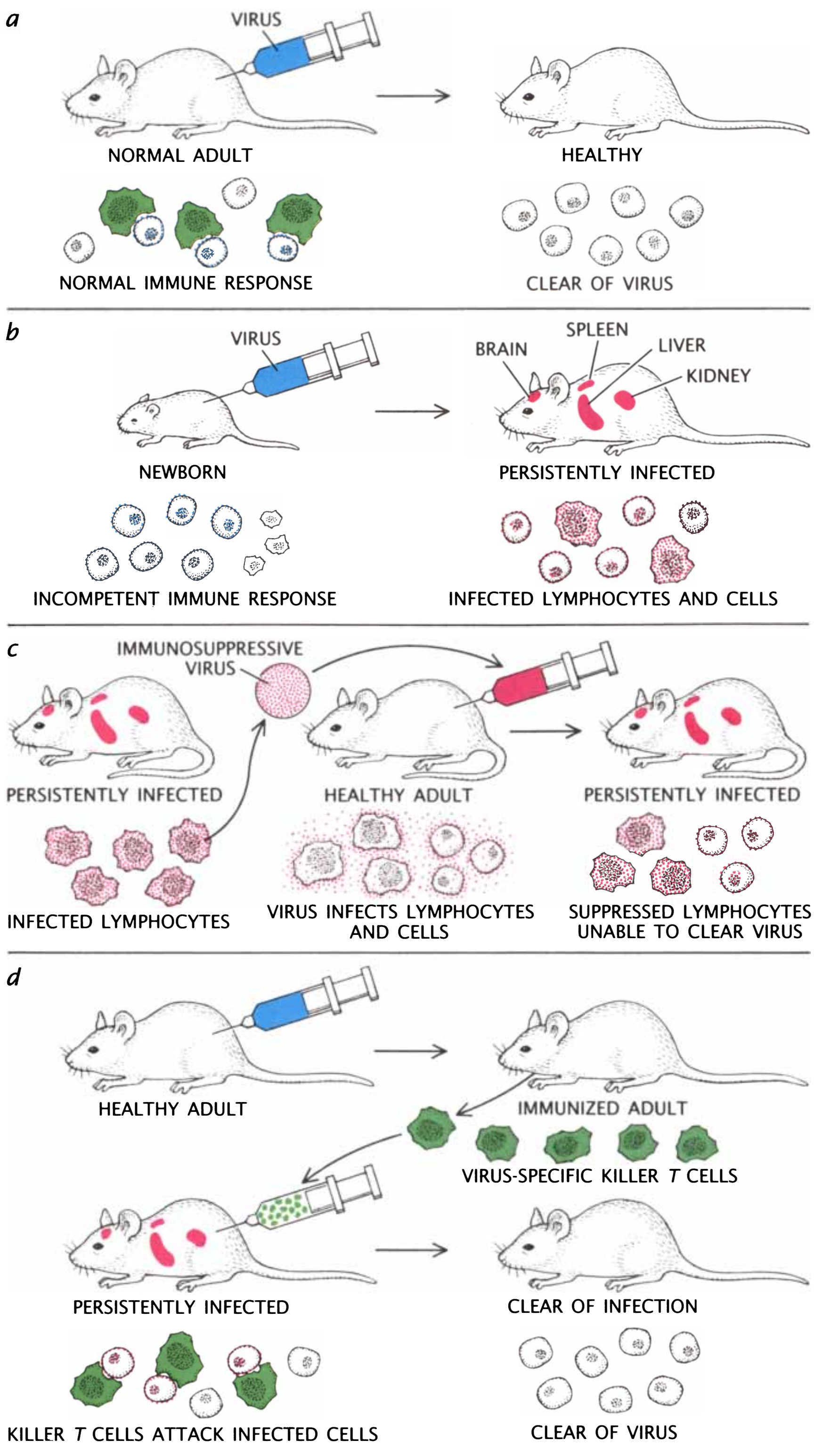


LCMV-INFECTED MOUSE (C3H strain) is compared with its uninfected littermate. The infected mouse's stunted growth rate and low blood glucose level were caused by a shortage of growth hormone. The shortage was traced to a sharp reduction in the initiation of transcription of the messenger RNA coding for growth hormone. The infection had no effect on the messenger RNA coding for other proteins, such as actin and thyroid-stimulating protein. The infected mouse regained normal growth and glucose metabolism after receiving a transplant of healthy, growth-hormone–producing pituitary cells.

could excise the genes from the Armstrong strain and reinsert them into the WE strain to make recombinant viral genes. Alternatively, we could make site-specific mutations in the genes to see if we could disable the disease-causing portion. Unfortunately, neither method can yet be applied to the particular class of viruses to which LCMV belongs. Instead we have been trying to compare directly the nucleotide sequence of the short-strand RNA in the disease-causing strain with the sequence in the benign strain. Such studies have not yet revealed a discrete site to account for the disease mechanism.

We had found a persistent viral infection that caused a hormonal deficiency without visible injury to infected cells. Was this a unique phenomenon or, rather, a general one involving diverse systems, such as the nervous system, the immune system and endocrine glands other than the pituitary? We quickly found that the phenomenon was not unique. In a different strain of mice, LCMV selectively invaded the beta cells of the islets of Langerhans in the pancreas. These cells normally secrete insulin, but in infected mice insulin production was apparently impaired, and the mice developed signs of diabetes. Yet the infected islet cells appeared normal under the microscope and showed no sign of inflammation. These studies showed that a virus could persistently infect islet cells and result in a biochemical and morphological picture comparable to that of adult-onset diabetes in humans.

We obtained similar results with thyroid epithelial cells and with neurons. Linda S. Klavinskis of Scripps found that LCMV established persistent infection in thyroid epithelial cells and perturbed the production of thyroid hormone. Although the cells showed no morphological damage, the blood levels of two thyroid hormones, T_3 and T_4, were reduced, as was the amount of messenger RNA encoding thyroglobulin, a component of both hormones. Similarly, W. Ian Lipkin of Scripps found that LCMV infected neurons containing the neuropeptide somatostatin but not those making cholecystokinin or gamma-aminobutyric acid (GABA). The infected neurons also showed no morphological injury, but they contained significantly less messenger RNA coding for somatostatin than did neurons from uninfected animals. The effect was specific to somatostatin: messenger RNA for cholecystokinin and GABA did not decrease in infected animals.



IMMUNE SUPPRESSION by LCMV was uncovered in the series of experiments depicted here. When LCMV (*blue*) was injected into normal adult mice, the animals remained healthy (*a*); infected cells were destroyed by killer *T* cells. When the virus was injected subcutaneously into newborn mice, the animals developed a persistent infection (*b*); not only were the immature lymphocytes unable to clear the virus but also a subset of the lymphocytes became infected. When healthy adults were injected with virus from the infected lymphocytes, they also became persistently infected (*c*); the virus had mutated into a strain (*red*) that infects and suppresses the lymphocytes. When uninfected, virus-specific killer *T* cells (*green*) from LCMV-immunized mice were injected into persistently infected mice, they cleared the virus (*d*).

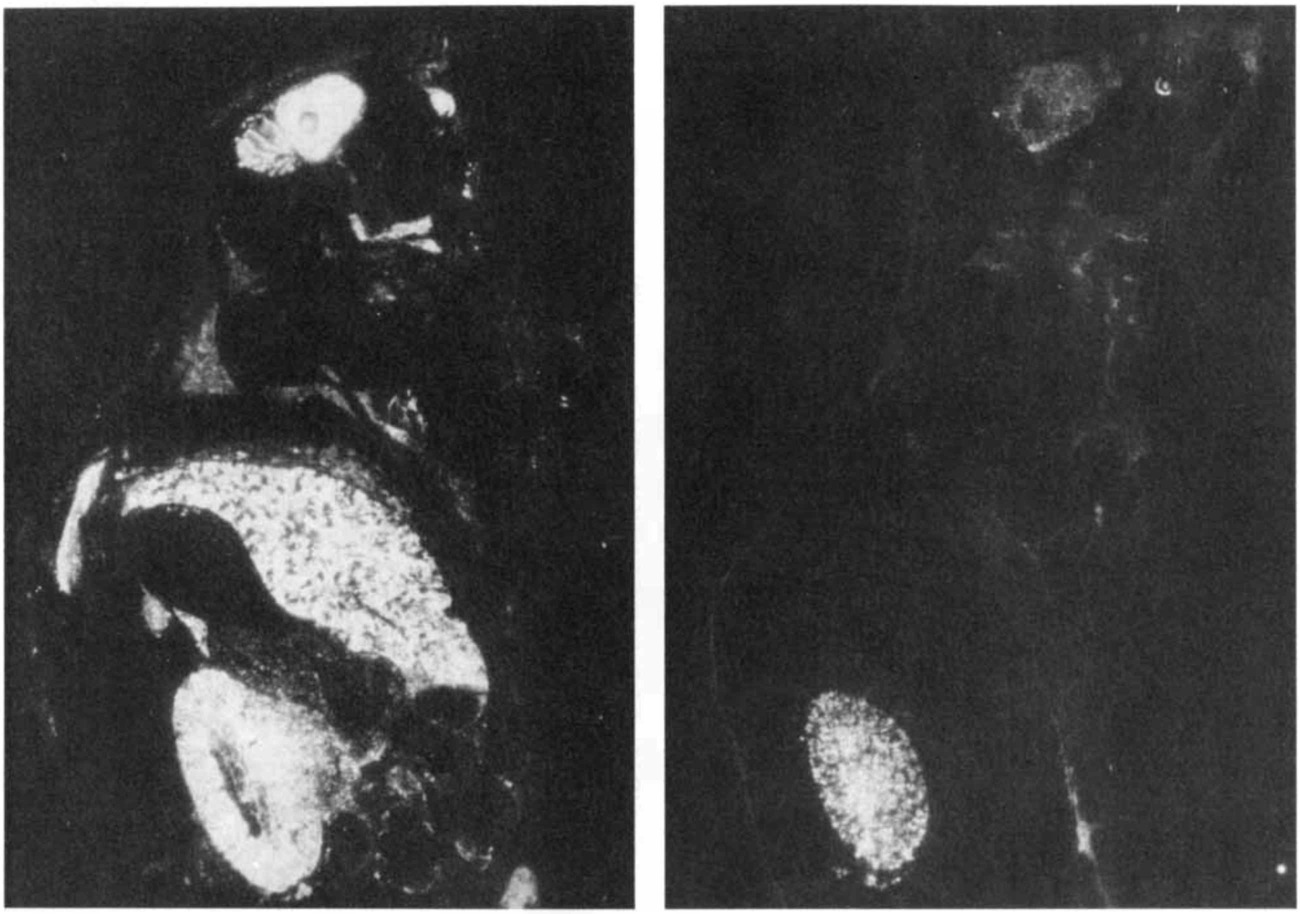
Other viruses have since been found to establish persistent infections in various organ systems, in humans as well as in animals, with analogous disruption of specialized cell functions.

In order to establish a persistent infection, viruses must be able to evade an organism's immune defenses. One common way is for the virus to infect lymphocytes and also macrophages (another type of immune cell, which digests injured cells and foreign bodies). The infection of these cells may play a role in the spread of infection within a host, in viral latency and persistence and in the transmission of viruses to uninfected individuals. In particular, we think that infection of lymphocytes may be responsible for the selective immune suppression that accompanies persistent viral infections.

Normally, when an animal has an acute LCMV infection, the viruses are attacked by virus-specific cytotoxic *T* cells, also known as killer *T* cells. The virus-specific killer *T* cells recognize certain viral particles that are bound to a unique "self" glycoprotein, called a major histocompatibility complex (MHC), on the surface of host cells. Each killer *T* cell recognizes and destroys only a specific viral antigen in combination with a specific MHC. For example, killer *T* cells specific to LCMV-infected cells of a specific MHC type will not kill the same MHC-type cells infected by different viruses; similarly, they will not kill cells of a different MHC type even if they are infected by LCMV. The killer *T* cells are primarily responsible for destroying infected cells, thereby terminating the infection.

In mice that have weak immune systems, however, the situation is markedly different. Persistent infections can be initiated by injecting LCMV into mice whose immune systems have been suppressed by such treatments as X-radiation and administration of immunosuppressive drugs. Similarly, if newborn mice are injected with the virus, their immature immune systems are unable to clear the virus, and the animals develop a lifelong persistent infection.

When we analyzed lymphocytes from persistently infected mice, we found infectious viruses and viral genes in a small percentage of the lymphocytes. Because these viruses were able to invade lymphocytes and apparently suppress the immune response, we thought they might be different from the seed virus with which we had inoculated the animals. Ahmed



HEAVY LOAD of viral protein shows up in an autoradiograph of a longitudinal section of a mouse with a persistent LCMV infection (*left*). The virus is concentrated in the brain, liver, spleen and kidney. A second persistently infected animal was injected with killer *T* cells from an LCMV-immunized mouse. After 21 days the animal was free of viral proteins (*right*) except in the brain, which was clear by 60 days, and the kidney, where complexes of virus bound to viral antibodies had lodged.

and I decided to explore the question. The results of our investigation were startling: we found that the virus recovered from the lymphocytes had the distinctive ability to initiate persistent infection in adult animals. Since these animals had intact immune systems, we concluded that the virus found in the lymphocytes must be a mutant strain able to suppress selectively the immune response targeted against it. We suspect that the virus, having first initiated a persistent infection in an animal with a weak immune system, then had the chance to replicate and produce mutants that could suppress the virus-specific killer *T* cells.

We also made another interesting finding, one that could have important implications for the understanding of immune suppression by viruses. When we searched for LCMV in the lymphocytes of persistently infected animals, we discovered that only about 2 percent of the lymphocytes—and mainly the so-called helper *T* cells—contained viral genes. Of the 2 percent, only one in 50 was infectious (that is, contained replicating virus). The vast majority of LCMV-infected lymphocytes, then, harbor incomplete, nonreplicating variants of the virus. This finding suggests that a virus need knock out only a small fraction of lymphocytes in order to cause selective immune suppression. What is more, whole, replicating viruses may not be necessary; incomplete viruses may be effective in disabling lymphocytes. This behavior of LCMV bears a striking resemblance to that of the human immunodeficiency virus (HIV), the pathogen implicated in AIDS.

What mutation in the seed virus enabled some of its progeny to suppress the immune system? To find out, Maria Salvato in our laboratory sequenced the nucleic acids in the viral RNA. She established that the mutations in the long RNA segment of the virus are restricted to the section encoding the viral polymerase (the enzyme that transcribes viral DNA) or to the section encoding "Z protein" (so called because it is a type of zinc finger, a regulatory protein that binds to nucleic acid), or to both. Independently, Ahmed, who is now at the University of California at Los Angeles, and his colleagues also located the immunosuppressive mutation on the large RNA segment by a different technique: reassorting the short and long segments of viral RNA from the immunosuppressive and the parental strains of LCMV.

We conclude from these experiments that variants of LCMV are spawned in lymphocytes and that they abort the production of the very killer *T* cells whose function it is to clear LCMV-infected cells. The variants thereby enable LCMV infection to become persistent, so that viral nucleic acids and proteins, as well as infectious viruses, accumulate in the blood and tissues. It is not yet known how other persistent or latent viruses suppress the immune response, but it seems possible that similar scenarios are played out in infections by such viruses as HIV, cytomegalovirus and hepatitis *B* virus, all of which are

VIRUS AND HOST	TARGET CELLS	SYNDROME
LCMV IN MICE	PITUITARY CELLS	GROWTH-HORMONE DEFICIENCY, GROWTH RETARDATION, HYPOGLYCEMIA
LCMV IN MICE	HELPER *T* CELLS, KILLER *T* CELLS	SELECTIVE IMMUNE SUPPRESSION
LCMV IN MICE	THYROID FOLLICULAR CELLS	HYPOTHYROIDISM
LCMV IN MICE	BETA CELLS IN PANCREAS	DIABETES
LCMV IN MICE	NEURONS	REDUCED SOMATOSTATIN
CANINE DISTEMPER VIRUS IN MICE	UNKNOWN	OBESITY, REDUCED NOREPINEPHRINE AND DOPAMINE
MINK-CELL FOCUS-INDUCING RECOMBINANT VIRUS IN MICE	UNKNOWN	DEFORMED WHISKERS
VENEZUELAN EQUINE ENCEPHALITIS IN HAMSTERS	BETA CELLS IN PANCREAS	DIABETES
SEMLIKI FOREST VIRUS IN MICE	NEURONS	REDUCED GAMMA-AMINOBUTYRIC ACID
RABIES IN MICE	NEURONS	IRREGULAR BRAIN WAVES
COXSACKIE *B* VIRUS (OR COXSACKIELIKE SEQUENCES) IN HUMANS	MYOCARDIAL CELLS	HEART DISEASE

PROVED LINKS between persistent viral infection and specific disorders are listed in this table. In all cases the virus was observed in vivo to interfere with the specialized, "luxury" functions of cells without disturbing their vital "housekeeping" functions.

known to infect and alter the functions of lymphocytes.

An urgent question in the treatment of persistent infections caused by virus-induced immune suppression is whether there is any hope of repairing the damage once the immune defenses have fallen. Our further investigations with LCMV indicate that such repair may be possible. We injected persistently infected mice with the appropriate type of virus-specific killer *T* cells, taken from the spleen of healthy animals (of the appropriate MHC type) that had been immunized against LCMV. The results were truly remarkable: the mice became virtually free of infection. The transplanted killer *T* cells removed both the infectious virus and the deposits of viral nucleic acids. Similar cytoimmunotherapy might be applied to terminate other persistent infections; in each case, however, one must first identify the immune cell that has been disabled.

Evidence is accumulating that situations similar to those we have described in mice may be occurring in humans. M. A. Preece and his colleagues at the University of London have found that some children who have had a rubella infection have retarded growth and depressed glucose metabolism. In the two boys who were studied, both abnormalities were corrected by replenishing their growth-hormone levels. Among the many afflictions of people who were exposed to rubella in utero is diabetes, which develops in nearly 20 to 30 percent of the cases. Coxsackievirus, cytomegalovirus and mumps virus can acutely infect and injure the beta cells in the islets of Langerhans, and mumps virus can also attack thyroid cells; whether such viruses can persist is not known. And many common childhood viruses, such as measles, chicken pox and mumps, persist in neurons in the central nervous system throughout life.

Persistent viruses may also turn out to be implicated in damage to vital organs. In the past two years, L. Anchard and his colleagues at the Charing Cross and Westminster Medical School in London and Heinz-Peter Schultheiss, Peter H. Hofschneider and their colleagues at the University of Munich have linked coxsackie or coxsackielike enteroviruses to dilated heart muscle (cardiomyopathy), a lethal condition that sometimes requires a heart transplant. They found that 30 to 50 percent of a group of more than 70 patients had viral nucleic-acid sequences in their heart muscle, whereas biopsied tissue from some 40 other patients suffering from other heart-muscle diseases did not.

The evidence gathered over the past decade has demonstrated a new way in which a virus can do harm. Because we now know that viruses can disrupt the production of hormones and neurotransmitters, we think persistent viral infections may play a hitherto unsuspected role in many human diseases. Indeed, it would seem prudent to evaluate any disorder of the specialized functions of nerve cells, endocrine glands and the immune system for possible infectious causes.

The number of such disorders is vast: it includes hormonal abnormalities such as adult-onset diabetes, neuropsychiatric disorders and autoimmune diseases such as systemic lupus erythematosus and multiple sclerosis. Whether these diseases will ultimately turn out to have viral origins is impossible to say. What is clear is that one can no longer rely on the received wisdom about viral pathology to make that determination.

FURTHER READING

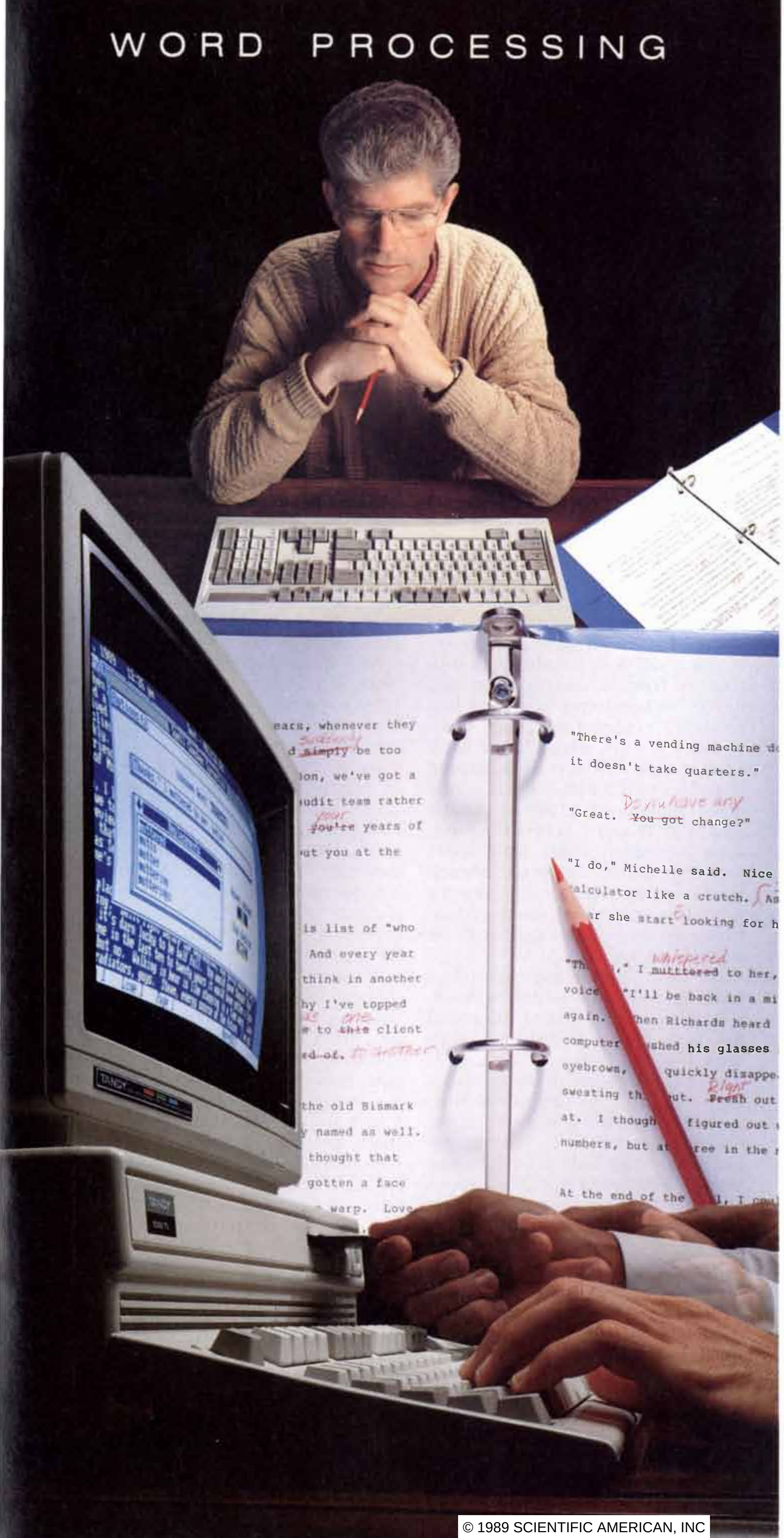
VIRUS CAN ALTER CELL FUNCTION WITHOUT CAUSING CELL PATHOLOGY. Michael B. A. Oldstone in *Concepts in Viral Pathogenesis*, edited by Abner Louis Notkins and Michael B. A. Oldstone. Springer-Verlag, 1984.

CYTOIMMUNOTHERAPY FOR PERSISTENT VIRUS INFECTION REVEALS A UNIQUE CLEARANCE PATTERN FROM THE CENTRAL NERVOUS SYSTEM. Michael B. A. Oldstone et al., in *Nature*, Vol. 321, No. 6067, pages 239-243; May 15, 1986.

VIRUSES PERTURB LYMPHOCYTE FUNCTIONS. M. B. McChesney and M. B. A. Oldstone in *Annual Review of Immunology*, Vol. 5, pages 279-304; 1987.

VIRAL PERSISTENCE. Michael B. A. Oldstone in *Cell*, Vol. 56, No. 4, pages 517-520; February 24, 1989.

LYMPHOCYTIC CHORIOMENINGITIS VIRUS SELECTIVELY ALTERS DIFFERENTIATION BUT NOT HOUSEKEEPING FUNCTIONS. Linda S. Klavinskis and Michael B. A. Oldstone in *Virology*, Vol. 168, No. 2, pages 232-235; 1989.



The Tandy® 1000

Runaway bestseller.

Whether you're writing a novel or a business proposal, you need to separate the facts from the fiction when it comes to word processing.

Fact is, the Tandy 1000 is the best-selling PC compatible in America. With good reason.

The Tandy 1000 comes with DeskMate software for word processing, filing—even graphics. So you can compose, edit, revise and polish your work as you see fit.

A built-in graphical user interface frees you from learning confusing computer jargon.

The Tandy 1000 will even proofread for spelling errors.

For more power, check out our line of business systems.

Because to writers everywhere, Tandy is a word to the wise.

Radio Shack®
The Technology Store℠
A DIVISION OF TANDY CORPORATION

The Mating of Tree Crickets

Their singing is a prelude to a fascinating array of reproductive strategies

by David H. Funk

At night, from late August until early October, old fields and woodlands across the U.S. reverberate with the melodious trill of crickets and katydids. To many human listeners the sound portends the passing of summer and the coming of fall; to others it provides a soothing reminder that the natural world is alive and well. To an entomologist the cacophonous singing of nocturnal insects is a signal that courtship, and with it the intense competition among males for females, is under way.

The most relentless and conspicuous of the nocturnal singers are members of the tree cricket subfamily Oecanthinae. Although they are in the same family (Gryllidae) as the more familiar black field crickets, tree crickets are smaller, ranging in size from 1.2 to 2.4 centimeters as adults, and they are pale green or yellowish in color. Moreover, as the name implies, they are mostly inhabitants of trees and tall shrubs, whereas field crickets are ground dwellers.

Of the 16 species in the U.S., the two that I have observed most closely are the broadwinged tree cricket *Oecanthus latipennis* and the two-spotted tree cricket *Neoxabea bipunctata;* both are common in the fields of southeastern Pennsylvania near where I live. Males and females of the two species mate repeatedly during the reproductive phase of their life, which generally lasts for from eight to 12 weeks, and an astute observer can readily spot courting males at night in the beam of a flashlight.

Males are easily recognized by the shape of their forewings, which are broader and more membranous than those of females. The overlying wing has an enlarged structural vein that acts as a file; when it moves across the underlying wing, which has a hardened margin that acts as a scraper, sound is produced [see "Cricket Auditory Communication," by Franz Huber and John Thorson; SCIENTIFIC AMERICAN, December, 1985]. Only males sing (as is the case for all crickets and most katydids), and their trill-like song, which is unique to each species, advertises a male's presence to sexually receptive females.

Receptive females approach males from the rear, because that is the direction in which most of the sound travels as it is amplified by and deflected off the wings. When a male senses the presence of an approaching female, he stops singing and turns around to touch her with his antennae. It is thought that by "tasting" her in this way he is able to confirm that she is a member of the same species and therefore an appropriate mate. The next stage in courtship involves an elaborate ritual (which varies slightly from species to species) during which the male again sings and exhibits other behavior such as rocking from side to side or pounding the substrate with his abdomen.

If the female is receptive to her suitor's overtures, she climbs onto his back and begins to feed on a special substance secreted from metathoracic glands near the top of his thorax, or wing-bearing section. The glands empty their secretions into a pit at the base of the male's hind wings, where they are readily accessible to the female once she is in the copulatory position. Although the secretions have not been chemically analyzed, they are highly attractive to females, and soon after a female begins to feed she will allow the male to mate with her.

Tree crickets package their sperm inside a hollow, nutrient-rich sphere called a spermatophore that is passed from the male to the female during copulation. Transfer takes only a few seconds, but for from five to 20 minutes afterward the female remains astride the male, feeding from his metathoracic secretions. As soon as she stops feeding and dismounts, the male resumes active courtship and will repeatedly attempt to back under her until she either remounts or wanders away. Such renewed activity on the part of the male may last an additional 20 minutes or more and is thought to be an important step in ensuring that large numbers of sperm pass from the spermatophore into a special sperm-storage organ in the female called the spermatheca. (When the female is ready to lay her eggs, sperm are released from the spermatheca and fertilize the eggs as they pass down the oviduct.)

The cricket spermatophore differs from the type found in most insects and other invertebrates in that much of it remains outside the female's

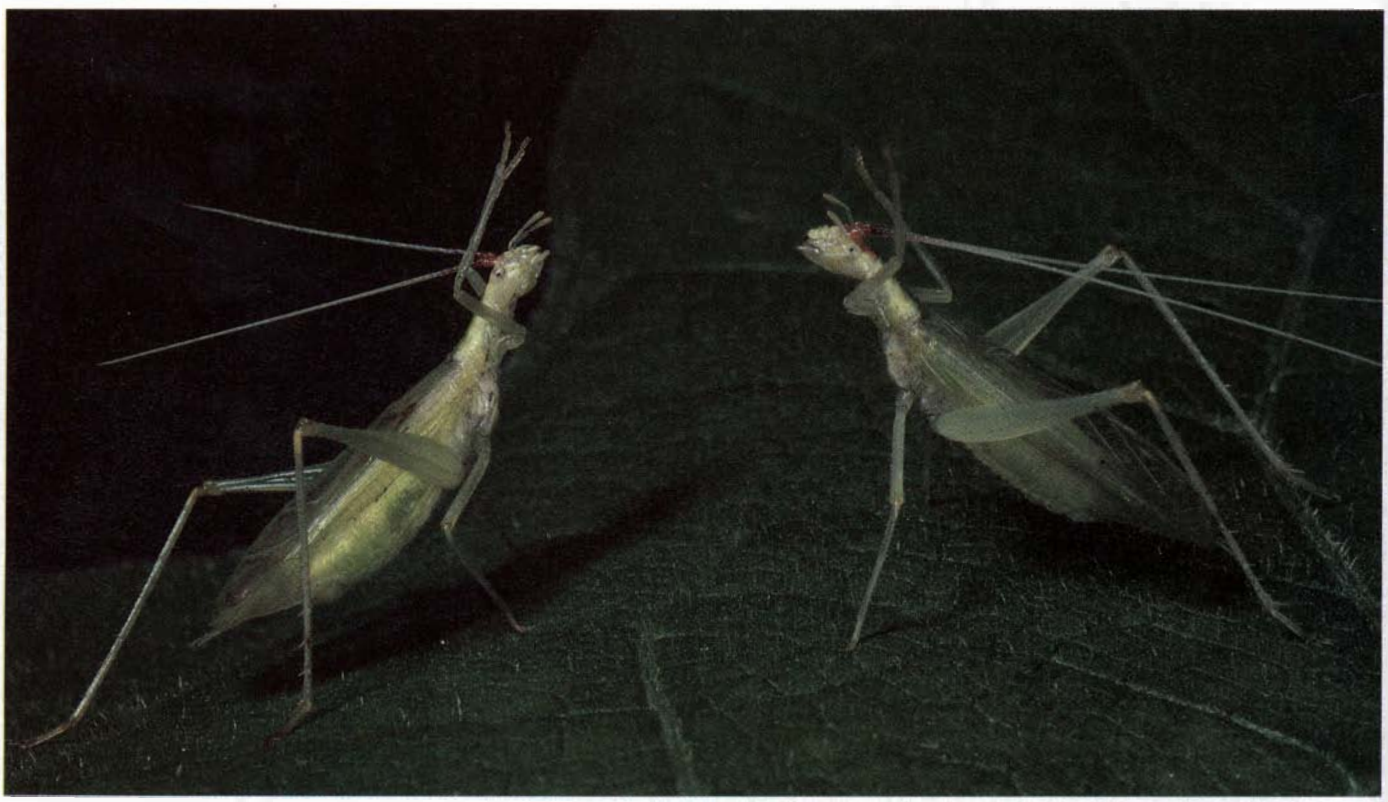
DAVID H. FUNK has been interested in insects since the third grade, when he began his first insect collection during a family outing to Georgia. Although he continues to collect insects, much of his leisure time for the past 10 years has been spent photographing and observing insects at night in fields near his home in southeastern Pennsylvania. Funk has a bachelor's degree in entomology from the University of Delaware and is an entomologist at the Stroud Water Research Center, an institution affiliated with the Academy of Natural Sciences of Philadelphia.

MALE TREE CRICKET* Oecanthus latipennis *sings at night from August through October and in doing so advertises his presence to nearby females. Sound is produced when the forewings are rubbed together; the resulting trill-like song is species specific.

TREE CRICKETS are sexually dimorphic: males (left) *have broad, membranous wings that they rub together to produce sound. Females* (right) *do not sing and have narrow wings that are held close to the abdomen.*

body after it has been transferred to her. Only a thin, threadlike tube enters the female; the ampulla, or main body of the spermatophore, dangles outside her genital opening. By curling her abdomen forward, a female can readily grasp the spermatophore with her mouthparts and remove it. Females always eat the spermatophore after copulation; they are thought to do so because the wall of the spermatophore contains nutrients that may enhance egg production. Males too will eat their own spermatophores (presumably for their nutritive value) if they are unable to transfer them to a female.

The glandular secretions and the external spermatophore of tree crickets make these animals interesting models for the study of sexual selection. First proposed by Charles Darwin in 1859 in his book *On the Origin of Species,* the theory of sexual selection (in contrast to the more general theory of natural selection) offers an explanation for why there are so many morphological and behavioral disparities between the sexes. As Darwin put it, sexual selection results from the constantly recurrent "struggle between the males for possession of the females." As a consequence of that struggle, natural selection has favored the evolution of traits, such as the antlers on deer or the red combs of roosters, that make a male physically competitive with other males or make



AGGRESSIVE ENCOUNTERS between males sometimes occur when two individuals are courting the same female. Their interaction is mostly a ritual in which both males wave their legs aggressively. Sometimes one male walks away, leaving the other to mate, but often the two fight until one literally kicks the other off the plant. (The female is not shown.)

him more attractive to females—or that, in many instances, have both effects.

The concept of sexual selection has been expanded and refined since it was first proposed. It is now commonly believed that many behavioral differences between males and females are related to the size of their gametes and the degree to which each parent invests resources in his or her offspring. As a rule, sperm, which are small, are energetically relatively cheap to produce (they contain only genetic material), whereas eggs (which are larger than sperm and contain nutritive as well as genetic material) are not. The ability of a male to produce offspring is therefore limited largely by his ability to inseminate as many females as possible; in general males can maximize their reproductive potential by mating with multiple females, and to do so they often compete—sometimes fiercely—with one another. Females, on the other hand, produce relatively few eggs, and because each egg represents a substantial energy investment, a female can maximize her reproductive potential by mating selectively.

Consequently, in many insect (and other animal) species, after an elaborate courtship ritual the male will part company with the female almost immediately after he has inseminated her. Such a strategy is not universal, however, and it is less likely in species in which males provide females with a resource (generally a territory, a nest site or a food item) and thus contribute in a significant way—other than genetically—to the reproductive output of the female. In fact, many tree crickets, including *O. latipennis* and *N. bipunctata,* do not separate immediately after copulation but instead stay

DURING COPULATION the female sits astride the male and allows him to link his genitalia with hers. The actual transfer of the sperm-filled spermatophore, which is not visible, takes only a few seconds. The male's wings are held vertically during and after copulation; in that position his metathoracic pit is exposed and is accessible to the female.

AFTER TRANSFER of the spermatophore (the main body of which can be seen dangling from the tip of the female's abdomen), the couple disengage their genitalia. The female remains with the male, however, for 20 minutes or more, feeding on secretions that are produced by metathoracic glands located at the base of her partner's hind wings.

METATHORACIC-GLAND SECRETIONS are highly attractive to females. They collect in a special pit (left) *at the base of the male's hind wings, where they are accessible to the female when she is positioned above the male. Crickets (as well as other insects) package their sperm in a special sac called a spermatophore* (right), *which is transferred to the female during copulation. Only the threadlike portion, through which the sperm pass, is inserted into the female; the ampulla, or main body (which is about from one to four millimeters long depending on the species), remains outside her body.*

together while the female continues to feed from the male's metathoracic glands. Why should males, who might otherwise pursue additional females, continue to feed their mates after the transfer of the spermatophore?

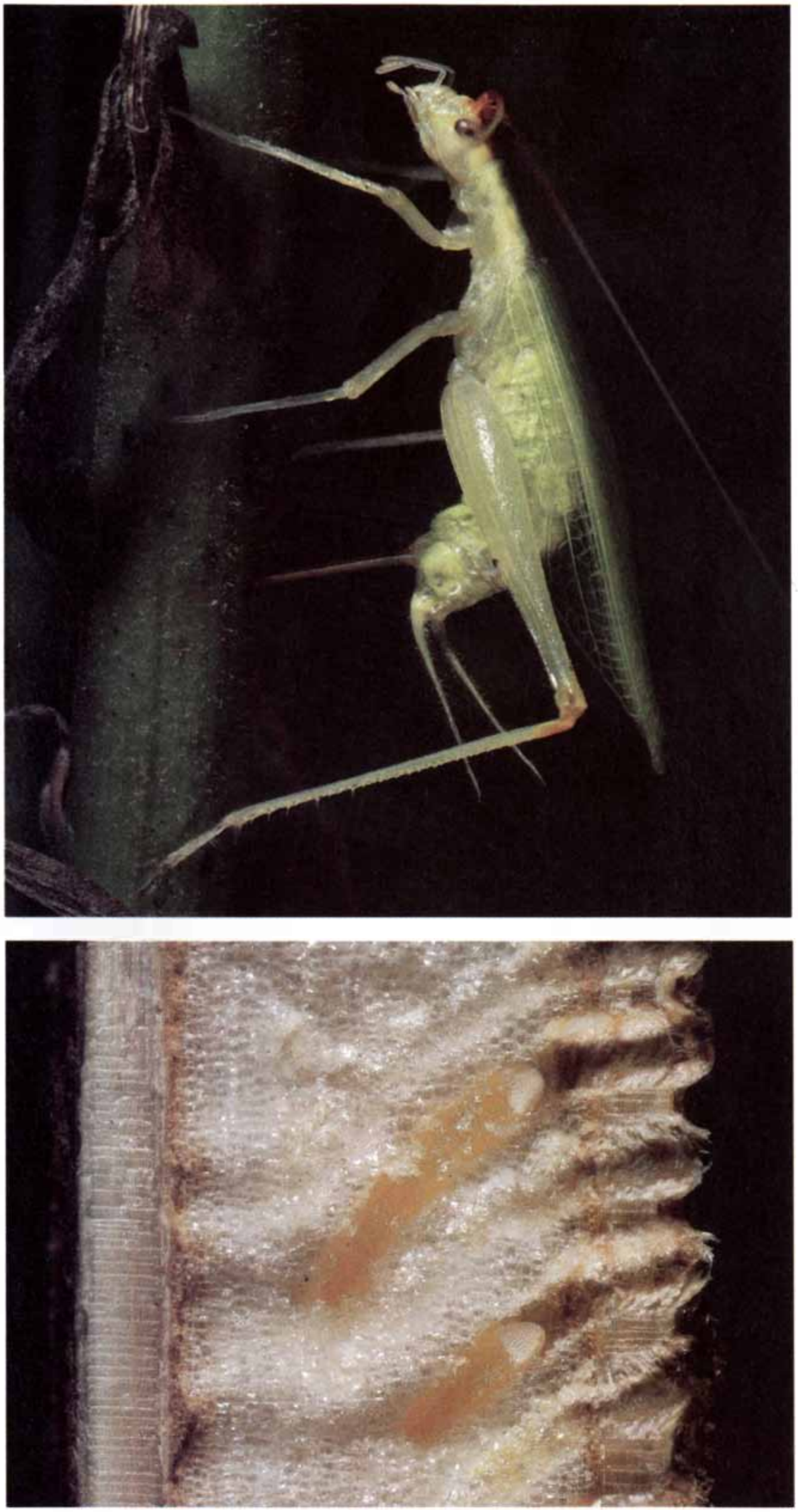
Although no studies have been done to show they contain valuable nutrients, the male's metathoracic secretions may contribute in some nutritive way to egg production. By providing his mate with nutrients that enable her to produce more eggs or larger ones, a male can effectively increase his own reproductive output. Even if the secretions do not enhance egg production, it is clear they serve another important function: they distract the female so she will not remove the spermatophore until it has been emptied.

If the female removes the spermatophore before all of the sperm pass out of it, the male has wasted part of his reproductive potential. In some cases a female who has mated previously, and therefore already has a sufficient supply of sperm in her spermatheca, may mate with a male simply to obtain the nutritive rewards of his spermatophore. The fact that many females tend to remove the spermatophore before it has been emptied of sperm supports that hypothesis.

From the male's point of view, of course, mating without insemination represents lost time and energy (in the form of spermatophores, glandular secretions and, in some species, body parts), and selection therefore strongly favors males who have strategies for ensuring successful insemination.

Males of *N. bipunctata,* for example, have evolved a secondary strategy for minimizing the premature departure of the female and thus maximizing their reproductive potential. In addition to secretions, they have evolved a behavior that prevents the female from leaving before most of the sperm have passed from the spermatophore into the spermatheca.

Males hang upside down from the leaves or twigs of trees when they are courting and will allow a female to mount them only while they are in that position. Once the female has mounted, the male drops his hind legs, maintaining his hold on the tree with his first two pairs of legs. That frees the hind legs for the next step: as soon as the spermatophore is transferred to the female, the male begins to shake rapidly while moving his hind legs back and forth along the body of the female. This behavior may persist for as long as 45 minutes, during which time the female is more or



***FEMALES DEPOSIT** their eggs in plant stems* **(top).** *The needlelike ovipositor (seen protruding from the abdomen) is inserted at almost a right angle into the stem's pithy interior. (The two appendages near the tip of the abdomen are cerci; they are sensory in function and occur in both sexes.) Females of some species lay their eggs in rows: the ovipositor is inserted, an egg is laid and the ovipositor is withdrawn and then reinserted a millimeter or so down the stem. Four holes and two eggs are visible* **(bottom)** *in a longitudinal section of a stem enlarged 20 diameters.*

Bill Rice, the bug, the plow, and the Lands' End Interlochen Knit Shirt

Bill Rice, ready to meet the President in his favorite Interlochen Knit Shirt.

There wasn't much moving on that shimmering hot summer day not long ago, on a West Texas ranch. Just a skittish lizard, here and there. And a tractor, plowing the fields, driven by a man named Bill Rice.

Now, summertime and bugs seem to go together like bacon and eggs. So it wasn't surprising that pretty soon, a bug crawled down into the Lands' End Interlochen Knit Shirt that Bill happened to be wearing that day.

(While we have gotten pretty well acquainted with Bill Rice, we can't tell you anything about this bug, whether it was a brave bug or an ornery bug, a wise old bug or a foolish young bug. It disappeared without giving any interviews.)

Anyway, we'll let Bill tell you what happened at this point, in his own words.

"There wasn't room for me and the bug, so I yanked the shirt off and just hung it on the back of my tractor.

"Well, to make a long story short, the shirt fell down and I accidentally plowed it under. After I plowed it back up and threw it in the washing machine, it was still as good as new. Now that's a well-made shirt."

A shirt good enough to meet the President.

At this point in the story, as Bill talked to us, he began to warm to the topic of our Interlochen Knit Shirts.

"I wear them everywhere I go. Working out here on the ranch, and going into town too."

Bill says it's the comfort of the shirt that's got him sold. Along with the way it looks. (He owns six of the shirts he's pictured in here!)

"If I was going to meet the President," says Bill with a twinkle in his blue eyes, "I'd wear this shirt."

What's so special about our Interlochen Knit?

As its name somewhat prosaically implies, it's knit with an "interlocking" stitch to create a fabric that is especially lofty and plush. Of pure combed cotton, no less, for surpassing softness next to the skin.

In one of our more exuberant moments, we once declared that our Interlochen Knit Shirt is absolutely "COTTON BALL SOFT!"

More unusual than the knit fabric itself is the fact that our Interlochen is even tailored with a soft touch. The collar is taped, the shoulder and armhole seams neatly single-needle stitched, the placket smoothly finished inside—all so no rough edges of fabric ever come between you and the softness of our Interlochen.

Why take such pains with a simple knit shirt? For the same reason we add extra features to our Oxford Shirts, our Corduroy Pants, our Canvas Attaches:

Quality is one of the ruling principles of The Lands' End Experience. And our customers seem to appreciate that.

Listen to Bill Rice again: "Out here, we bust our butts to make a living. And when we spend money, we want quality. That's why I buy from Lands' End."

There's more to The Lands' End Experience.

To make sure our kind of Quality is accessible to as many people as possible, we insist that it always be accompanied by Value. As Direct Merchants, we deal directly with our manufacturers, sell directly to our customers, to deliver the utmost for the dollar.

Completing The Lands' End Experience is our efficient, enthusiastic brand of Service.

We ship most orders within 24 hours after we hear from you. Back everything we sell with an emphatic guarantee that reads, simply, GUARANTEED. PERIOD.® And keep our "store" open 24 hours a day, seven days a week.

Why not call one of our operators today, toll-free at 1-800-356-4444, and let them introduce you to The Lands' End Experience? They're famous for their friendliness.

So friendly are they that sometimes, our customers are encouraged to tell them the wildest stories you've ever heard. Like the one, not long ago, about the Texas farmer who plowed under one of our knit shirts.

Now, who'd believe that?

LANDS' END® DIRECT MERCHANTS

Lands' End, Inc. Dept. Q-22
Dodgeville, WI 53595

Name ______________________

Address ______________________

City ______________________

State ____________ Zip ____________

Call toll-free:

1-800-356-4444

TWO-SPOTTED TREE CRICKET Neoxabea bipunctata *has an unusual mating strategy: couples hang upside down, and the male moves his hind legs back and forth along the body of the female. The movement of the legs is thought to prevent the female from leaving the male before most of his sperm have passed from the spermatophore into her body.*

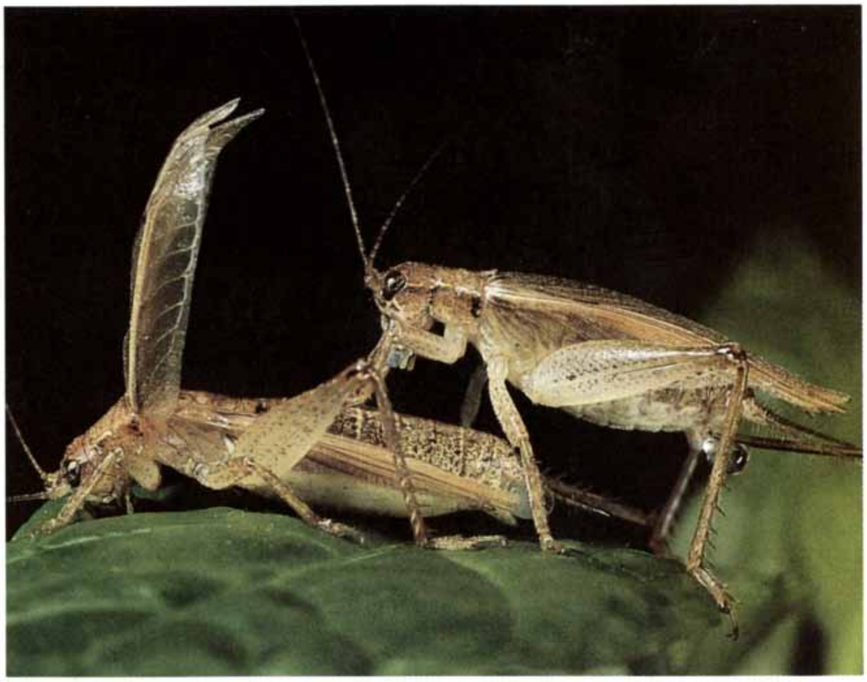
AFTER THE FEMALE leaves the male, she curls her abdomen forward and removes the spermatophore with her mouth. She then eats it. The spermatophore, which is now emptied of most of its contents, is thought to be a source of nutrients.

less trapped by the living fence her mate has created around her. Mating pairs of two-spotted crickets are highly conspicuous during this phase and can be easily located with a flashlight even when they are high in a tree. Eventually the male stops moving, and the female dismounts, whereupon she removes the spermatophore and eats it.

Other cricket species have evolved different strategies for preventing females from leaving before the spermatophore is empty. Richard D. Alexander of the University of Michigan and Daniel Otte of the Academy of Natural Sciences of Philadelphia have found, for example, that males of the restless bush cricket *Hapithus agitator* (subfamily Eneopterinae) retain the attention of a female by letting her nibble at their forewings for 10 minutes or so during copulation.

Males of another eneopterine cricket, *Orocharis saltator,* produce a long series of spermatophores—as many as 20—during a single sexual encounter. The spermatophores differ from those of many crickets in that each contains relatively few sperm (about a tenth as many as are in a typical *Hapithus* spermatophore).

Sperm transfer in *Orocharis* takes place in assembly-line fashion: immediately after receiving her first spermatophore, the female will back away from her partner and in doing so will drag her abdomen across a leaf, dislodging the spermatophore, which she then begins to eat. A minute or two later the male, who has produced a second spermatophore, will attempt a second mating. If the female is receptive to his overtures, she climbs onto his back where she continues to eat the first spermatophore while being inseminated with a second. When she has finished eating, she again dismounts, removing the second spermatophore, which she also eats. Meanwhile the male, who has produced a third spermatophore, initiates yet another coupling. The entire sequence is repeated many times: I have seen mating pairs remain together for as long as three and a half hours, during which time the male produces a continuous flow of spermatophores. Because it takes about nine minutes for a female to consume one spermatophore but only about four minutes for sperm to empty from the spermatophore, it appears that sequential spermatophore production in *Orocharis* serves to distract the female in much the same way that metathoracic secretions do in *Oecanthus.*



SPERM TRANSFER in the bush cricket,* Orocharis saltator, *resembles an assembly line. The first spermatophore was removed by the female soon after she received it and is now being eaten by her. After she removed the first spermatophore, the male produced a second one, which he has transferred to the female, who is still eating the first one. Meanwhile the male has produced a third spermatophore, which he will give to the female as soon as she removes (and begins to eat) the second one.

Sexual selection is clearly responsible for the diverse array of mating strategies seen in crickets. Beyond that, many questions remain. Perhaps the most interesting is the extent to which males contribute nutritionally to egg production. Resources such as secretions and spermatophores help to ensure insemination, but do they also lead to the production of more or healthier offspring?

Studies by Darryl T. Gwynne of the University of Toronto indicate that in some katydid species the male produces an enormous spermatophore representing as much as 40 percent of his body weight. The spermatophore is far larger than is needed for insemination: a large portion of it consists of a nutritive covering that contributes to egg production. In other species, however, the spermatophore has only a small nutritive covering. Scott K. Sakaluk of Illinois State University has found, for example, that in one species of cricket the spermatophore covering is only large enough to distract the female during insemination.

One possible explanation for this disparity is that the size of a male's contribution reflects the number of partners a female is likely to have during her reproductive life. If the number is high, as it is in many crickets, one male's nutritional resources may end up supporting a competitor's offspring. Therefore, in species where multiple matings are common, selection should favor males who provide resources that are no larger than are necessary to ensure insemination. Although such modest contributions are unlikely to satisfy a female's nutritional demands, males are unlikely to compromise their ability to secure additional mates by investing in offspring of dubious paternity.

FURTHER READING

THE EVOLUTION OF INSECT MATING SYSTEMS. Randy Thornhill and John Alcock. Harvard University Press, 1983.

ORTHOPTERAN MATING SYSTEMS: SEXUAL COMPETITION IN A DIVERSE GROUP OF INSECTS. Edited by Darryl T. Gwynne and Glenn K. Morris. Westview Press, 1983.

MALE CRICKETS FEED FEMALES TO ENSURE COMPLETE SPERM TRANSFER. Scott K. Sakaluk in *Science,* Vol. 223, No. 4636, pages 609-610; February 10, 1984.

THE CONTRIBUTION OF MULTIPLE MATING AND SPERMATOPHORE CONSUMPTION TO THE LIFETIME REPRODUCTIVE SUCCESS OF FEMALE FIELD CRICKETS (*GRYLLUS BIMACULATUS*). L. W. Simmons in *Ecological Entomology,* Vol. 13, No. 1, pages 57-69; February, 1988.

COURTSHIP FEEDING IN KATYDIDS BENEFITS THE MATING MALE'S OFFSPRING. Darryl T. Gwynne in *Behavioral Ecology and Sociobiology,* Vol. 23, No. 6, pages 373-377; 1988.

SCIENCE AND BUSINESS

Tussling over technology, growing diamond films, microchip futures, the jobless

A Technological Fix

The U.S. searches for a stand on technology

Summers in Washington, D.C., can be torpid affairs; the weather weighs heavily on all, and Congress takes a holiday. This summer, however, may see some surprises. Industries from superconductor makers to computer companies are increasingly looking to the government for financial and legislative support in their fight to stay abreast of international competitors. The government, dogged by fiscal constraints, is trying to sort out its response.

The project that has become something of a test case is high-definition television (HDTV). According to proponents, who include leaders of such companies as AT&T, Motorola and Zenith, HDTV is more than a better picture on the tube. They argue that HDTV display technology will be integral to advanced computers. HDTV sets will also be ravenous consumers of memory chips, thereby creating a strong market for domestic producers. All in all, the industrialists argue, a vibrant domestic HDTV business could revitalize U.S. high technology.

But then comes a noticeable change in the pitch. Business leaders also maintain that the cost of developing an HDTV industry is too great for them to bear alone. Borrowing capital to build manufacturing plants is expensive, and the expected profits from HDTV are small, at least for the first decade or so. As a result companies interested in HDTV have banded together under the banner of the American Electronics Association to lobby the government.

An HDTV proposal forwarded by the AEA to Congress in early May has several features reminiscent of other industry-government proposals. Industry will provide the people and develop the technology while the government helps out with funding and low-cost loans—more than $1 billion over five years, the plan suggests. Changes in antitrust laws to protect a manufacturing consortium from antitrust charges would also be necessary.

Both the administration and Congress have created special task forces for sorting out the government's HDTV stance. The Department of Commerce was slated to release a plan for government action on July 1. Significant funding for HDTV, however, seemed unlikely to be part of the plans. Financial "supplements for industry," through either direct funding or tax credits, look unrealistic in this "era of fiscal limits," according to Representative Frederick Boucher of Virginia. Instead Boucher suggests that antitrust amendments are a more likely first step. Congress is already entertaining at least five bills that prescribe different procedures for gaining exemptions from antitrust damages. One bill, co-sponsored by Boucher and Representative Don Edwards of California, would exempt approved ventures from antitrust damages.

Other federal efforts to foster technology are also running aground on budget issues. Late last year Congress voted to reshape the National Bureau of Standards into a new agency, the National Institute of Standards and Technology, which would broadly support technology. But the 1990 budget proposed for NIST was several million dollars less than the previous NBS budget. "Right now NIST is severely handicapped by the lack of funds," says Russell C. Drew, a past president of the Institute of Electrical and Electronics Engineers. Moreover, as of early June, the White House had not yet nominated either an undersecretary for technology or a director for NIST.

The Defense Advanced Research Projects Agency has tried to fill the civilian technology void: DARPA earmarked $30 million for HDTV research and awarded the first five contracts in June. The agency is spending another $15 million on superconductivity research and $100 million for SEMATECH, the semiconductor-manufacturing research consortium. But the defense budget has also been slimmed down, leaving the defense community less than eager to fund what many believe are primarily civilian projects.

Senator John Glenn of Ohio is hoping to tackle that problem by introducing a bill this summer that would create a civilian twin to DARPA, the Advanced Civilian Technology Agency. Last year Glenn's proposal to reorganize the Department of Commerce into a department of industry and technology met with skepticism; instead Congress adopted the more modest proposal of creating NIST. Glenn now hopes to argue successfully that a smaller agency with a larger budget built alongside of NIST could effectively fund civilian technology research. Although Drew supports the measure, he worries that Glenn's proposal will be stonewalled. Too few in Congress see the connection between competitiveness and funding for technology, Drew says.

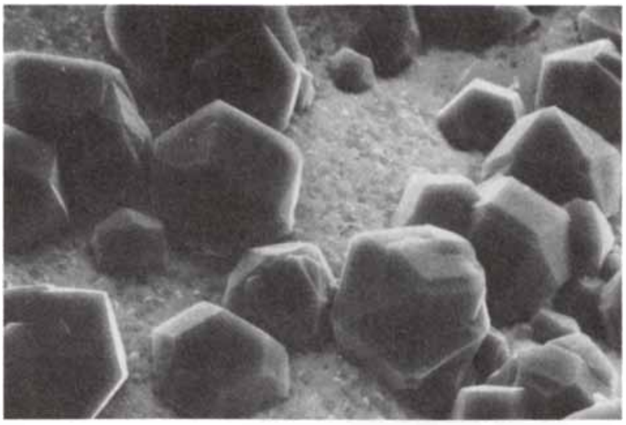
Some industry leaders, meanwhile, have realized that government assistance is not enough to restore a competitive edge. At a workshop at the National Academy of Sciences, Gordon Moore, chairman of Intel, reflected on the history of the U.S. semiconductor industry by observing that "each of us optimizes our company's situation, and probably for a moderately short-term [period], even if we think we are fairly long-term oriented." While those decisions may have helped the companies, Moore added, they have damaged the overall health of U.S. industries. —*Elizabeth Corcoran*

Don't Try Tiffany's

Thin diamond films capture the heart of industry

General Electric has worked with diamond for years; more recently, a scientific advisor challenged workers at General Motors to make thin diamond films. At Martin Marietta the manufacturing divisions requested that the company start a diamond-film research program.

Particularly in Japan, and more recently in the U.S., companies are racing to grow thin diamond films, lured by the promise of diverse applications—for example, as coatings to make tools and bearing surfaces resistant to heat and corrosion, as extra-hardened windows, as heat sinks for electronic devices and eventually



DIAMOND CRYSTALS (right), *10 microns in size, make up the thin "films" grown by researchers. In one technique for forming the diamonds* (left), *a hot tungsten filament excites a mixture of methane and hydrogen gas, releasing carbon atoms that are then deposited on a nickel substrate. Photographs courtesy of the Naval Research Laboratory.*

as high-speed electronic components. "In the near term, diamonds may be more important than high-temperature superconductors," says William F. Banholzer, a manager at the General Electric Research and Development Center.

Why? Pure diamond boasts an alluring collection of properties. An exceptional number of very strong bonds linking the carbon atoms that make up diamond make it the hardest material known. Diamond conducts heat five times faster than copper at room temperature and has a friction coefficient similar to that of Teflon. It is, moreover, virtually transparent to wavelengths from the ultraviolet to the far infrared, carries sound at high velocities and is resistant to most chemicals.

Yet its hardness also makes diamond difficult to shape to the needs of industry. Since the 1950's General Electric and other companies have been forging synthetic diamond by exposing graphite to high temperatures and pressures; the resulting material is best used for abrasives and cutting tools. In the 1970's Soviet researchers were among the first to grow thin diamond films under low pressures and on substrates other than diamond [see "The Synthesis of Diamond at Low Pressure," by B. V. Derjaguin and D. B. Fedoseev; SCIENTIFIC AMERICAN, November, 1975]. U.S. interest in diamond films quickened in the mid-1980's following vigorous diamond-film programs initiated in Japan.

A variety of chemical vapor deposition techniques are now used to create diamond films. Typically, a mixture of methane and hydrogen is converted to a plasma above a substrate such as silicon or nickel by heating the gas with a hot filament, microwave radiation or even an oxygen-acetylene torch, notes James E. Butler, a research chemist at the Naval Research Laboratory. The carbon in the gas condenses onto the relatively cooler substrate; the more reactive forms of carbon (such as graphite) combine with atomic hydrogen and are removed, which leaves only diamond crystals.

Yet the process of diamond growth is poorly understood, says Thomas A. Perry, a scientist at General Motors Research. Efforts to speed the deposition rate often result in more crystal defects and impurities. Although companies hesitate to reveal deposition speeds, most cite one micron an hour as typical; the NRL has formed "good-quality, clear crystals" at 50 microns an hour, and the Tokyo Institute of Technology has reportedly achieved rates as high as 930 microns an hour.

Workers are also looking for ways to lower the temperature of the substrate from 900 to 100 or 200 degrees Celsius; at such temperatures a diamond film could be deposited on plastic, for instance. Diamond-clad plastic could be used for everything from scratch-proof eyeglasses to reinforced airplane windows, Butler says.

Nevertheless, no one has yet grown the diamond likely to be most useful to the electronics industry: a film of single-crystal diamond, several centimeters across, on a nondiamond substrate. Since diamond is both a good thermal conductor and a good semiconductor, diamond microchips could serve as exceptionally high-speed devices. "Electronics will be the big money-maker," predicts Russell Messier, who heads an industry-university program on thin diamond films at Pennsylvania State University. But it will take about a decade to produce such devices, he cautions.

U.S. companies are therefore emphasizing basic research. Although Perry expects that General Motors and its subsidiary, Hughes Aircraft, might eventually employ diamond films as hard coatings and optical devices, GM's current program focuses on diamond chemistry, he says. General Electric's first thin-film applications "will probably be abrasives, because that's our business," Banholzer says; at this point GE workers are trying to understand what contributes to diamond's properties. Martin Marietta's work is still a "seed program," says Stephen R. Winzer, an associate director; he expects it might coalesce into a larger effort in a few years "depending on how well it fits with Martin Marietta."

In the meantime, diamond films are beginning to appear as niche products. Sumitomo Electric is providing at least two Japanese stereo equipment makers with diamond films for tweeters. Crystallume, a start-up company in Menlo Park, Calif., is selling diamond windows for scientific instruments; the windows are said to transmit low-energy X rays and will not leak pressurized gases. Nevertheless, while "the U.S. is on a par with the Japanese in the science," Butler says, "we're not on a par in the transition to immediate devices." —*E.C.*

Futuristic Markets

When is a computer chip like a pork belly?

Commodities traders have never been averse to swapping contracts based on the future prices of all sorts of bulk, standardized goods. Pork bellies, oil, gold, stock indexes, even tulip bulbs have stoked their business. Now two separate exchanges have proposed putting traders into the high-technology arena by creating futures markets for dynamic random-access memory chips.

DRAM's are the guts of computers and products such as laser printers and facsimile machines. The chips are produced in huge volume, primarily by Japanese manufacturers. One-megabit chips, for instance, which hold a million elements, top the current best-seller list.

Might DRAM's be considered a commodity, easily tradeable and fungible? Most certainly, according to the newly established Twin Cities Board of Trade, based in Minneapolis, and the Pacific Stock Exchange in San Francisco. In late May both groups announced their intention to start trading DRAM futures next year, pending regulatory approval. "It sounds like a neat idea," says Andrew Davis, a private commodities trader in Chicago. Chip makers and users, on the other hand, have greeted the suggestion with some hostility and much curiosity.

"To call a DRAM a commodity isn't right," protests a spokesman for Texas Instruments, one of the three U.S. companies that makes DRAM's. "There are different [chip] speeds, the packaging requirements are different for different [computer] manufacturers and so on," he says. "We think the idea is outrageous," confided an industry player who declined to be identified.

Commodity specialists respond that they already cope with great diversity in agricultural products. The different products require contracts that spell out their characteristics.

For buyers and sellers of goods, futures can help to insure against price changes, says William P. Tai, a vice-president at Memory Clearing Corporation in San Francisco, which is developing the DRAM futures market with the Pacific exchange. Large price swings are common in the chip business. According to the market research firm Dataquest, U.S. prices for one-megabit DRAM's were as high as $18.92 each in the third quarter of 1988 and are expected to drop to $16.53 in the third quarter of 1989. The changes exact a toll: this year Apple Computer's second-quarter earnings dropped significantly, in part because the company had bought one-megabit DRAM's at high prices. (Apple has not commented on the proposed DRAM futures market.)

Now consider the fortunes of a chip user who buys futures, Tai says. He might know in January that his company will probably need 10,000 one-megabit chips in July but may worry about a price increase. He buys a futures contract at $18 a chip; by July, prices have risen by $.50 a chip. The chip user can then sell his futures position, earning an extra $.50 a chip that will cover the higher costs of buying chips in the cash market. Or he can simply take delivery on the futures contract. Either way, the chips are no more expensive than he expected.

Chip makers such as Texas Instruments say that they believe in building "strategic, long-term relationships" with users to dampen the swings in prices and supplies. As a result Tai expects that the early players in the DRAM futures market will be companies too small to form such ties with the large manufacturers.

COTTON COMMODITIES EXCHANGE in New Orleans was portrayed by Edgar Degas in 1873. Could it be the forerunner of a futures market in microchips? Picture courtesy of the Granger Collection.

Both the Twin Cities and Pacific exchanges have long roads ahead, nonetheless. Twin Cities, which does not yet trade anything, says it will first introduce a cash market in DRAM's and then a futures market if its efforts are successful. (In a cash market, buyers purchase goods for immediate delivery.) Twin Cities is sponsoring its own electronic network to provide a data base of current information to prospective chip buyers, sellers and eventually speculators. The Pacific exchange, the major securities and options exchange in the western U.S., is proposing only a futures market.

Precisely what chips the exchanges decide to trade and how broadly they will write the terms of the contracts could make or break the DRAM futures market, observers say. Tai says companies have suggested a host of features they would want to see in a contract. Yet will there be enough trading volume to support a narrowly drawn contract for, say, a one-megabit static-column DRAM with 100-nanosecond access time, in a plastic package with four input/output pins?

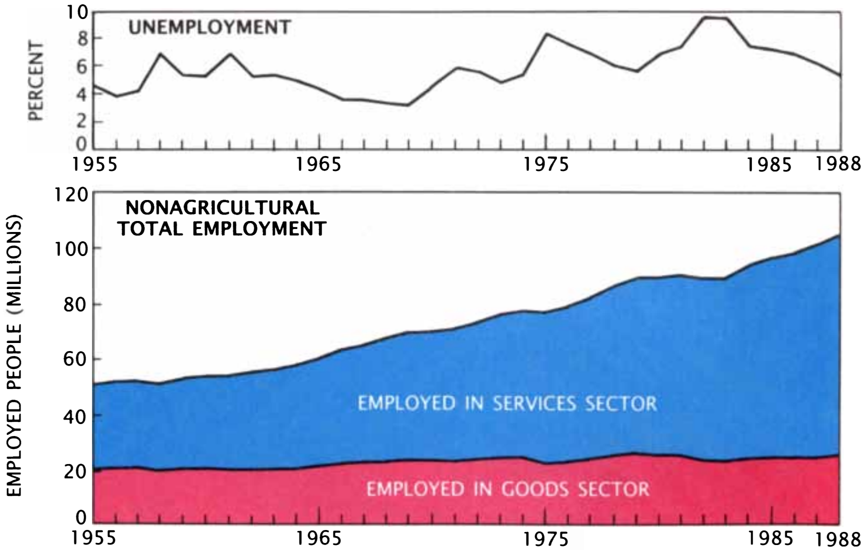
"I don't know if [a DRAM cash or futures market] is good or bad," says J. Richard Iverson, president of the American Electronics Association. "But I'd like to see it tried." —*E.C.*

THE ANALYTICAL ECONOMIST

Tracking the unemployed

At 8:30 A.M. on the first or second Friday of every month, Janet L. Norwood, commissioner of the Bureau of Labor Statistics, announces the latest calculations of the unemployment rate and number of employed. Citizens, business people and economists all await her report eagerly—but then interpret the numbers very differently. Some are cheered when the pool of the jobless has shrunk; others are worried that lower unemployment will signal inflation.

What is most clear about unemployment statistics is that they constitute a slippery and politically charged issue. Scrutinized on a monthly basis, small bumps in the unemployment rate swell to the dimensions of the Himalayas. To the discomfort of the financial markets, jittery at the faintest hint of inflation, economists wrangle over how low unemployment can fall before it sparks wage competition among employers. And underlying the reported numbers are trends



UNEMPLOYMENT OF THE CIVILIAN LABOR FORCE* (top) *has waxed and waned since the 1950's; the current rate is somewhat greater than 5 percent. Employment, based on a survey of company payrolls* (bottom), *has grown slightly in the goods-producing sector* (red); *growth has been stronger in the services-producing sector* (blue). *Data from the Bureau of Labor Statistics.

that have dramatically different effects on the country's various demographic and geographic groups. Ironic though it may seem, unemployment says little about how people are faring.

Consider the newspaper reports trumpeting that the unemployment level had fallen a tenth of a percent to 5.1 percent in May and that nonagricultural employment had risen by 101,000 jobs to a total of 108.2 million.

In fact, the Bureau of Labor Statistics is the first to point out that the fluctuations amount to statistical noise. A change in monthly unemployment rates must exceed .2 percent to be statistically relevant, and changes in employment must involve more than 140,000 jobs. Financial analysts trying to predict inflation are interested only in changes involving more than 250,000 jobs, according to Thomas J. Plewes, an associate commissioner at the BLS. Even so, dancing on data points is still in vogue; when the *New York Times* asked three noted economists about the significance of May changes, one predicted wage increases would quicken, another implied wage boosts would lessen and the third said that at this point unemployment would not really affect wages.

There is another reason to look at the numbers skeptically: not everyone out of work is "unemployed." To be counted as part of the labor force, individuals have to be actively looking for work and ready to start immediately. The BLS estimates there are approximately 855,000 "discouraged" workers who have given up hope of finding a job; they are not counted among the unemployed. Part-time workers, on the other hand, are counted as employed, even if they are looking for a full-time job.

For some, unemployment takes on an uglier shape than a national rate of 5.1 percent would suggest. Black men and women between the ages of 16 and 19 continue to suffer the highest unemployment rates; in March 32.8 percent were looking for work. (Unemployment among white teenagers, in contrast, was 11.9 percent.) Unemployment also has a geographic bias: parts of Texas have jobless rates exceeding 18 percent, whereas areas of Connecticut have barely 2 percent.

Still, on the average, the story of the past few years has been one of declining unemployment and rising participation in the work force, "against the drumbeat of economists who claimed that [the rates] couldn't come down," says Barbara R. Bergmann of American University. During days of higher unemployment, some economists argued that there was a so-called "natural rate" of unemployment, below which the economy risked the woes of high inflation. As more women, minorities and "baby-boomers" swelled the labor force, economists predicted that the natural unemployment rate would rise.

Such arguments, however, were little more than post hoc efforts to explain away rising jobless rates, observes Lawrence H. Summers of Harvard University; "the natural rate has had an alarming tendency to follow the real rate." Unemployment among women, for example, is now about the same as that among men.

Most of the new jobs have come in the services-producing sector, a predictable and healthy trend, Bergmann says. The services sector may weather business cycles better than goods-producing industries do, according to Jonathan S. Leonard of the University of California at Berkeley. "If unemployment goes up, people will still go to the bank and [still] get their hair cut." On the downside, he believes that individual jobs may be less stable.

The additional jobs were also created without boosting real wages. Since 1973 the average real wage (that is, a paycheck minus inflation) has stayed the same. Yet blended into that average was a decline in the real wages of workers with less than a high school diploma and a rise in the wages of college graduates, reports Sar A. Levitan of George Washington University.

Barry Bluestone of the University of Massachusetts at Boston argues that this split is widening faster in the services sector than it is in the goods sector: "Just think of the professionals who work in the New York high rises during the day and the janitors who come in at night." In 1987, he says, college-educated professionals were making at least 3.52 times as much as those who lacked high school diplomas. (In the goods sector, college graduates made only 2.42 times more than high school dropouts.) In the face of foreign competition based on inexpensive labor, Summers adds, the outlook for U.S. laborers without high school diplomas is likely to get worse.

Bad news? Or a healthy transition? After all, more young people are going to college and taking those higher-paying jobs, says Marvin H. Kosters of the American Enterprise Institute. Even Bluestone's numbers concur: in 1987, 19.5 percent of the labor force had college degrees, up from 10.3 percent in 1963. Moreover, only 18 percent had less than a high school diploma, down from 45 percent in 1963.

Still, it seems that to some extent the U.S. work force is being split into two camps. In one are the college-educated, who are enjoying rising incomes; in the other are high school dropouts, whose share of the economy is dwindling. How far apart the groups might drift depends on whom you ask. In any event, unemployment data are unlikely to give a hint.

—Elizabeth Corcoran and Paul Wallich

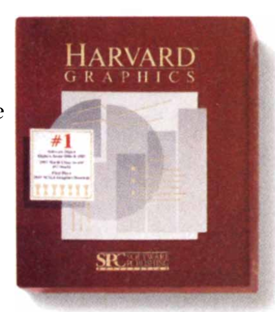
After Lunch, Have Your Audience Eating Out Of Your Hand.

Now you can turn a bright idea over lunch into a feast for the eyes in no time at all. With Harvard Graphics™ from Software Publishing Corporation.

Just pick the graphics format you want from the menu. Then enter your data, hit a key and your chart or graph appears on screen in an instant.

Customizing your graphics can be just as easy. Convert a flat pie chart to 3-D in one stroke. Combine charts or display data tables. Spell-check any length presentation in just one step. Then produce your polished business graphics on paper, slides or overheads. For meetings in boardrooms or lunchrooms.

So call today for our free trial disk. You'll see the qualities that make Harvard Graphics the top-selling, most award-winning program of its kind. With it, you'll be serving up presentations no audience can resist.



FOR A FREE TRIAL DISK CALL 1-800-345-2888 OPERATOR 53

SPC SOFTWARE PUBLISHING CORPORATION

Runs on IBM PC and MS DOS compatibles. © 1989 Software Publishing Corporation, 1901 Landings Drive, Mountain View, CA 94039-7210

The Metamorphosis of Information Management

Information refineries convert facts into knowledge. They rely increasingly on the power and elegance of parallel programming

by David Gelernter

If Gustave Eiffel were alive today, he would be working in software. What iron, steel and reinforced concrete were in the late 19th and early 20th centuries, software is now: the preeminent medium for building new and visionary structures. The continuing progress of computer hardware technology has given software builders new sources of power, and new principles of software design allow them to harness that power for new kinds of computation.

On the high plateau of modern computer hardware, software systems are arising that are more than fast data processors or glorified adding machines—they are information refineries that can transform mere facts into knowledge on a vast scale. For instance, information-refining programs might transform the low-level data that describe a hospital patient, transportation network or factory into a high-level synopsis. They might convert electronic file cabinets full of data into authoritative discussions of the objects or events (patient histories, wildflowers or automobile accidents) contained in those files.

DAVID GELERNTER is associate professor of computer science at Yale University. The need for additional computing power to support sophisticated programs led to his interest in parallelism. He and his colleagues have developed a parallel-programming system called Linda that they believe is suitable for building complex software. Now Gelernter is focusing again on specific parallel applications. He received his undergraduate education at Yale and his doctorate in computer science from the State University of New York at Stony Brook. Gelernter's article "Programming for Advanced Computing," appeared in SCIENTIFIC AMERICAN in October, 1987.

The development of information refineries is intimately bound up with the growing role of parallelism in computer science. Parallel hardware consisting of multiple subcomputers in a single box has increased the computing power available to users. Parallel software makes it possible to tap this power, and frequently it offers elegant solutions to complex information-management problems.

A computer program is a kind of machine: this fact is a good starting point for understanding the importance of parallel software. A program is a machine for transforming information, just as a stamping press, for example, is a machine for transforming materials—put steel (or information) of one shape in, get steel of another shape out. Programs are ordinarily defined as lists of instructions that tell a computer what to do, but this definition is no more helpful than defining a book as ink and paper. Software machines are composed of the same stuff they transform: information encoded as numbers and characters. They rely on physical computers to decode the information and carry out the requisite actions.

The idea of programs as machines is a key to new principles of software design that rely on parallel instead of sequential execution. Conventional programming techniques limit programs to one action at a time. Most physical machines, however, carry out many operations in parallel; a stereo system, for example, simultaneously spins a compact disk, detects its digital coding, converts digital signals to analog, amplifies analog signals and converts electricity to sound. Parallel software machines may sound exotic, but they are more like "real" machines than conventional software is.

Parallel software machines, together with powerful parallel computers to run them, open the prospect of new kinds of programs performing tasks heretofore undone. For all that computers have accomplished in the past 40 years in the way of calculations and data processing, they still offer relatively little help in managing information—refining numbers or signals into useful knowledge.

We and investigators elsewhere have been exploring a number of approaches to building such information refineries. Two kinds of machine hold promise: the information filter, which transforms an incoming stream of data into higher-level knowledge, and the "smart" data base, which sorts out interesting patterns from records of many similar objects or events.

A "transportation watcher" is an obvious example of information filtering. Such a machine would gather the millions of small facts that describe a transportation system at a given minute and convert them into usable information tailored to the needs of any traveler. It could answer such questions as: What's the best route to work this morning? What's the best route across the city for a long-haul trucker? Where should a tow truck be dispatched? Which bridge is most likely to collapse next? Software like this exists now only on a small scale.

A similar machine could filter the flood of data generated by sensors and attendants in an intensive-care unit in an attempt to find disturbing patterns before they became crises. Any complex, fast-changing system could be monitored in the same way—be it a car, a factory, an oil tanker or a national economy.

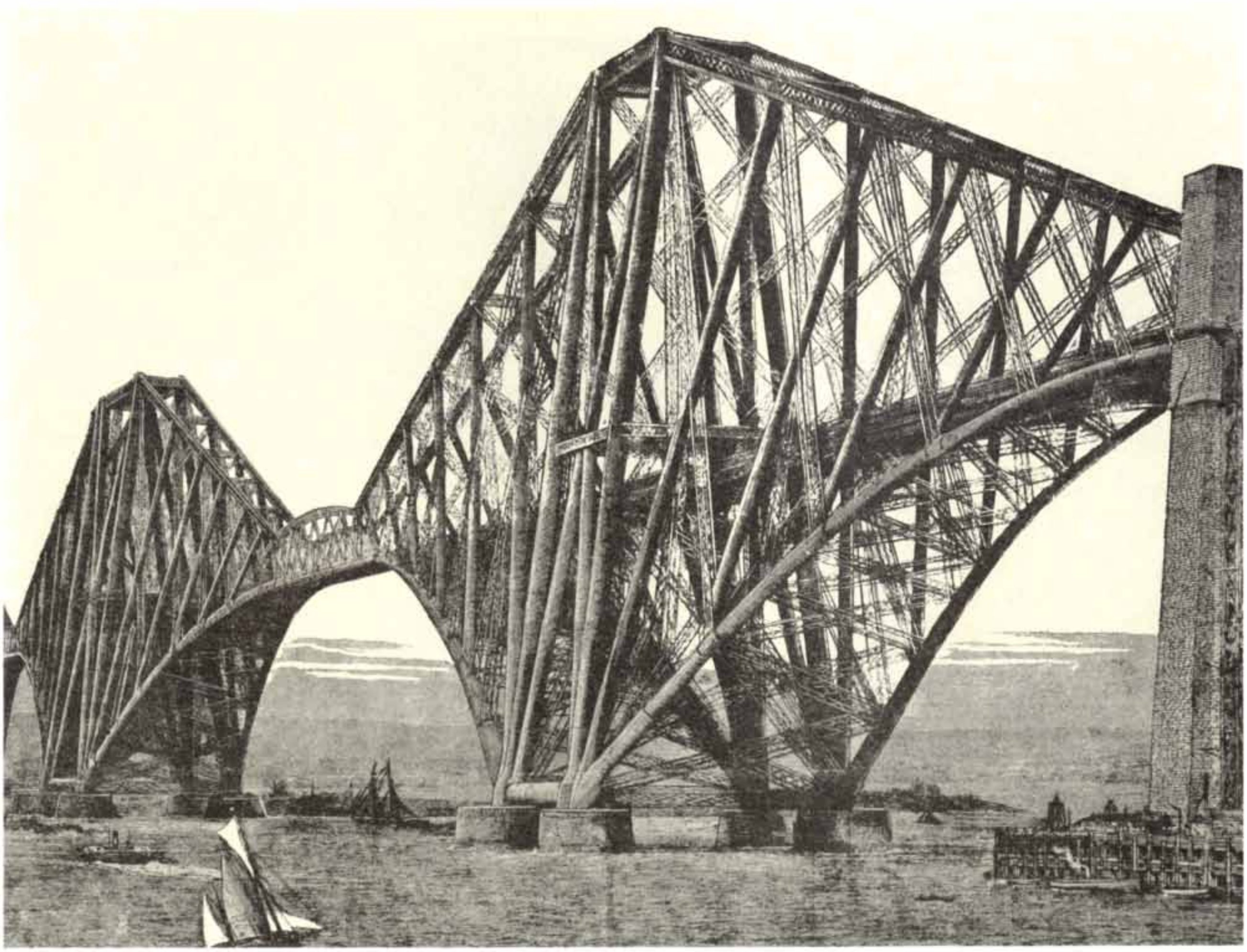
A smart data base would do its work on stores of existing data. It would take a description of some problematic object or event—a chest X ray, a patient's case history or some institution's financial and business status—and elucidate the situation on the basis of outcomes recorded for thousands or perhaps millions of similar cases. In the course of doing so, the data base would pull from its files exactly those records especially pertinent to the case at hand. A smart library drawing on the latest news from around the world might take up temporary residence on anyone's desktop, expecting to be asked not "Retrieve all items containing the following list of keywords ..." but rather "Tell me about Oliver North." Only the simplest precursors of such systems exist today. In data bases consisting of news articles, technical literature and many other kinds of material, data that could be converted into valuable information are being underused or simply wasted.

Even if sophisticated information-management machines were available, most computers are too slow to run them effectively: computers now can process large data sets in simple ways (generating millions of telephone bills) or small amounts of data in sophisticated ways (generating the image of a page for a laser printer). Processing large volumes of data in sophisticated ways might take weeks or months.

The increased computing power offered by parallel hardware offers some help, but a more important problem remains: building sophisticated software for information management is an inordinately complex task. Few organizations have the time and money, and even fewer the expertise, to do so. Most of the projects undertaken so far are at universities or other research institutions that lack the resources to turn their prototypes into finished, widely used systems. A Department of Public Software Works may ultimately be needed to build the massive information refineries the U.S. could use.

Building parallel software of any kind is a new field (as is software construction in general). Competing approaches differ in what they demand of the programmer and in what they offer. The most conservative approach is to build "parallelizing" compilers: programs that would take ordinary (nonparallel) software and convert it automatically to parallel form. This approach is attractive because it allows old programs to be upgraded with minimal effort, but it works only for certain kinds of software and can achieve only a limited parallel speedup. Most computer scientists agree that for computers with many processors automatic translation software cannot compete with real programmers.

Another approach extends more or



VISIONARY ENGINEERING of the 19th century produced structures of cast iron or steel such as the Forth Bridge near Edinburgh. Today the leading medium for building visionary structures is software. Although construction material has changed, many of the guiding principles remain the same: complex systems must be built out of many relatively simple parts, and the detailed design of the individual parts should not affect the details of other parts or of the structure of the whole.

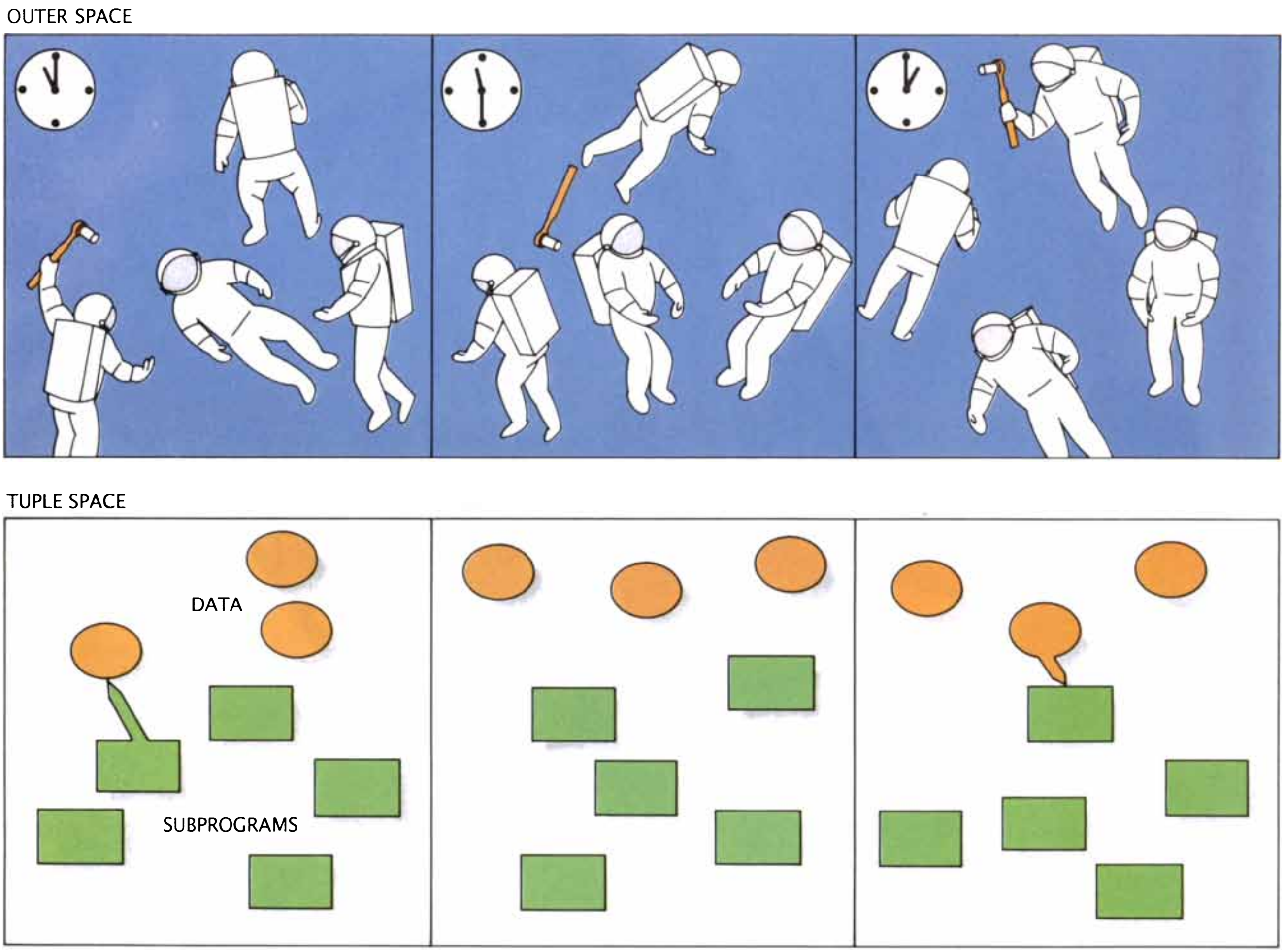
less conventional programming languages so that some operations can be performed simultaneously on many pieces of data. All the elements in a list can be updated at the same time, for example, or all the items in a data base scanned simultaneously to see if they match some criterion. This approach lets programmers replace iteration (repeated execution of the same set of instructions with different data) with parallel execution. It does not address a more general case, however: performing many interrelated but *different* operations at the same time. This ability is essential in developing complex information-management machines.

Our group at Yale has developed a parallel-programming system, called Linda, that addresses the problem of coordination in general—how to build big programs out of many concurrently active smaller programs no matter what they are and in what language they may be written. In Linda, coordination and computation are two separate issues of equal standing that together define the problem of building a software machine. Linda is still considered a somewhat heterodox (if not subversive) idea in computer science, but the number of sites where Linda is used is growing steadily.

Linda's approach to coordination is to simplify as much as possible. Rather than setting up intricate communication paths among the parts of a software machine and complex protocols by which two or more parts can synchronize their actions, Linda makes communication anonymous and uncoupled. A component that produces data need not know who will use it or when. Components that require data need not know who produces it.

A good way to understand Linda-style coordination is to imagine a crowd of astronauts building a space station. An astronaut who has finished using a wrench lets go of it—sets it adrift. Another astronaut who needs the wrench can reach out and grab it. The same goes for any other tool, for a list of tasks that need doing or for some piece of information that might be shared by several astronauts. Whoever produces the information simply releases it where anyone who needs it can look at it. Individual astronauts don't know who has something they need or who needs what they have; they just pick up tools or set them afloat. Nor do they synchronize their actions: a wrench set adrift by one astronaut at 10 o'clock might be picked up by another three hours later.

The hypothetical astronauts inhabit physical space; Linda programs inhabit what we call "tuple space." (A tuple is a chunk of data; the term is a generalization of terms such as quadruple and quintuple.) Passive tuples are just data available for reading or processing. Active tuples are subprograms, all



COMMUNICATION between cooperating subprograms in the Linda parallel-programming model is anonymous and uncoupled. Chunks of data called tuples are created by one subprogram and can be read or manipulated by any other subprogram at a later time. Linda-style communication is analogous to that among imaginary astronauts building a structure in orbit: an astronaut who has finished using a tool sets it afloat where anyone else who needs it can pick it up. Neither the giver nor the receiver needs to know the other's identity. Linda was developed by the author and Nicholas Carriero, also of Yale.

executing simultaneously, that consume and produce other tuples. Active tuples turn into passive ones, available for reading or processing, once they have finished executing.

Linda and other parallel-programming systems provide the tools for building large software machines and the raw materials for construction. What might the architecture of an information refinery look like? One possibility is the "blackboard" machine designed by Lee D. Erman, Victor R. Lessing and their colleagues in the course of pioneering work on speech recognition at Carnegie-Mellon University. The blackboard architecture employs any number of modules, each of which writes its results in a shared storage area for other modules to use. Information refineries can also be based on the assembly of simple rules into complex deductive chains, a powerful technique developed by Bruce G. Buchanan, Edward A. Feigenbaum, Edward H. Shortliffe and their colleagues at Stanford University.

An architecture we find promising has the form of a trellis: a row of modules at the bottom connected to sensors in the real world, a second, higher row to refine the data and make connections between different items, a third row for further refining and so on. Two-way communication between rows permits the lower-level modules to alter their actions in response to queries or comments passed down from upper levels.

We have built a prototype trellis machine to monitor and analyze patient data in an intensive-care unit. This research system, developed in collaboration with Perry L. Miller of Yale's department of anesthesiology and medical informatics, should help us develop and fine-tune techniques for filtering and refining information. (The current version was built largely by Michael Factor, with major contributions from Aaron I. Cohn and Dean F. Sittig.) The bottom-level software modules in the trellis are designed to be connected to machines that monitor heart rate, temperature at various points, blood pressure at various points and so on; higher-level modules focus on increasingly more general questions about the patient's condition. Modules directly above the bottom rank look for trends or obvious noise (erroneous readings) in the data; the modules above them look for simple patterns, and still higher modules look for diagnoses that might involve the presence or absence of many simple symptoms.

Modules in the upper reaches of the machine assess the likelihood that some complex pattern or condition holds true. For example, in the current system an assessment of the likelihood of hypovolemia (low fluid volume) depends on an assessment of "clamping down," which depends on systemic vascular resistance, which in turn depends on blood pressure (a datum derived by applying a procedure called Kalman filtering to raw blood-pressure data from a sensor).

Metaphorically, each higher-level module has a gauge whose dial runs from "definitely false" to "definitely true," plus a special region marked "insufficient data." (So far we have had encouraging results in building the modules so that they weigh the incoming information as a clinician would, but getting the probabilities right is an ongoing research effort.)

Data flow upward from module to module through the connections in the trellis. When new data percolating up from below cause a particular module to reassess its view of the patient's situation, the module sends its new assessment upward, and higher-level modules reassess their viewpoints in turn. Users can query a module that has not formed a judgment, causing the module to send a request for more information down through its connections. If the additional information can be found (or generated), it will be dispatched back up the trellis.

The trellis machine is an inherently parallel piece of software. All of its modules run continuously and concurrently because it would make no sense for them to do otherwise. Starting up one module, running it, stopping it, running another and so on would be needlessly complicated. Parallelism is also important for information-filtering machines such as the patient monitor because these systems must respond quickly despite a rapid influx of data and the substantial amount of analysis that may be needed. Parallel computers provide the power needed to run the parallel software machine we have built.

The trellis is distinguished from other architectures by two main characteristics: it is transparent and locally comprehensible. The trellis's transparency means that it is easy to figure out what each part of the machine does; the structure of the connections between modules reflects the structure of the problem the machine is designed to solve.

This quality makes it easy for the user to understand what the trellis can do, what it cannot do and how to use it effectively. Transparency is an important (but often neglected) issue for the design of any machine, and it is crucial for a complex machine used under high-pressure conditions. From a display that shows the status and interconnections of the trellis elements, the user might rapidly determine, for example, that the machine has ideas about two forms of congestive heart failure (but none about kidney disease) and that the judgments about heart failure depend on information about ischemia, cardiac output and so on.

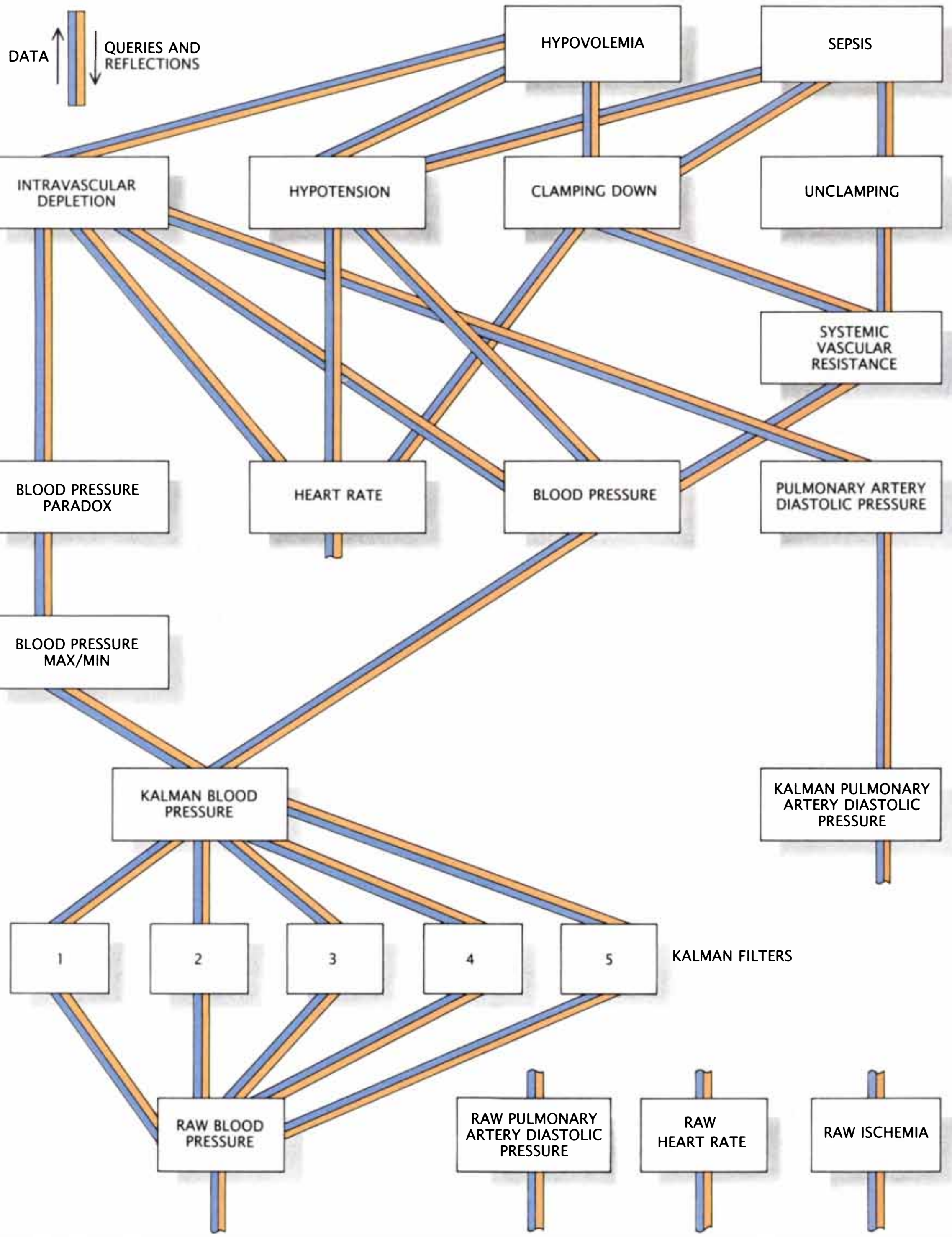
Local comprehensibility means that understanding any single module in the trellis requires understanding only its immediate neighborhood, not the trellis as a whole. Programmers who change modules or add modules to the trellis need understand only the modules to which their new ones will be directly connected. A trellis incorporating thousands or even tens of thousands of modules could be designed, assembled, tested and put into service, even though no single programmer would understand the machine in its entirety.

When it is completed, the ICU monitor that we have been developing will probably incorporate several hundred modules. Other trellises, such as the hypothetical transportation watcher, might be far larger. From the lowest-level modules monitoring conditions on particular sections of road, railbed or runways on up to the top-level modules responding to sophisticated questions about transportation alternatives, the transportation trellis machine could comprise a million nodes or more. Data would stream upward constantly, while perhaps thousands of requests for new data to clarify the ramifications of existing information filtered downward.

Whatever its problem area, a trellis machine addresses the same basic question—"What's going on here?"—at many levels of detail. Installed in the face of onrushing streams of data, trellises transform otherwise wasted data into information.

Information filters are powered by flowing data. Other kinds of information refineries draw on the pools of information accumulated in large data bases. High-speed parallel data bases could inaugurate the same kinds of changes in working with data as the advent of railroads introduced for travel. Rail lines cut the travel time from, say, North Cholmondeley to London from five days to five hours. They introduced no fundamental change: a trip that had always been possible merely became faster and cheaper. Railroads changed the way people thought about travel because what had previously been a serious undertaking became routine.

Many data bases would be much more useful if they could be searched quickly. For example, when a new DNA sequence (perhaps from a new virus) is discovered, geneticists may need to know what previously catalogued sequences it resembles. Answering this question may require a fair amount of computation to determine the similarity between the new sequence and each sequence in the catalogue. The larger the data base, the more time-consuming and expensive the search. Gene data bases are already large, and making them larger is an important



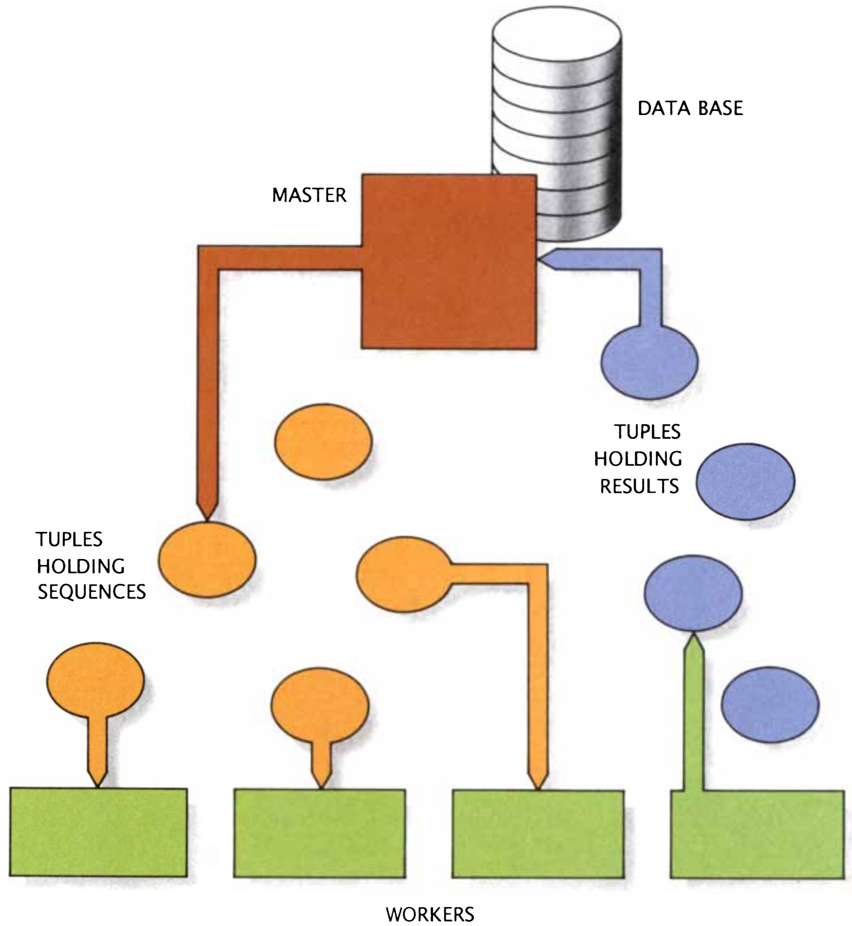
TRELLIS MACHINE developed by the author and his colleagues monitors a patient in an intensive-care unit. Sensors at the bottom send information (*blue*) up the hierarchy to units that consider ever more complex hypotheses about the patient's status (arrows indicate connections between units). Higher-level units may send queries and comments (*orange*) down the trellis to elicit additional information or to change the behavior of a lower-level unit based on a high-level hypothesis.

goal of research in molecular biology. This problem is by no means specific to biology. There are many other examples, such as searching catalogues of chemical reactions, identifying images by comparing them with libraries of examples or retrieving data files on the basis of complex search criteria.

Our group has built a prototype mass-search machine for the DNA-sequencing problem out of two types of software modules: one master to parcel out the data for analysis and any number of workers to perform the comparisons. In essence, the master gives each worker a copy of the search target (the new sequence) and then reads through the data base, grabbing each sequence and dropping it into tuple space. Each worker module takes an unexamined sequence out of tuple space, compares it with the target and drops the result back into tuple space. The master module gathers all the results and presents a summary to the user when the search is complete. (The details of the process are a little more complicated; for example, under some circumstances the master module will focus many workers on a single large comparison rather than leaving it to a single worker.)

To test the concept, we ran the mass-search machine on a 64-processor parallel computer at Yale. Each processor is powered by an Intel 80386 chip that is roughly comparable in speed to the fastest of current-generation personal computers. One of these processors would take just under four hours to compare a median-length target sequence against the primate section of the GenBank sequence data base. The mass-search machine accomplishes the same task in just over four minutes. This speedup confirms that we can do an hour's worth of computing in a minute by focusing about 60 processors efficiently on the same problem.

The parallel computer we used for this test is large and expensive, but the mass-search machine should perform just as well, proportionally, on smaller desktop parallel computers. Even more interesting, the mass-search program could run on a "virtual parallel computer" consisting of many unrelated machines linked by a network (like the networks that currently transmit electronic mail and transfer files from one location to another). Networks are the Ultima Thule of parallel computing; only a few researchers have ventured there. Many programs that thrive on conventional parallel computers will wither and die on networks. The mass-search program, however, appears to be well suited to network execution because it has a relatively low ratio of communication to computation.



MASS-SEARCH MACHINE matches an unknown DNA sequence against known sequences in a data base. The master module creates tuples containing the known sequences; worker modules compare them with the target sequence and generate tuples containing the results of the comparisons. The master module then collects the tuples to determine which catalogued sequences the target most closely resembles.

During the summer of 1988 Robert Whiteside of Sandia National Laboratories in Albuquerque ran a series of Linda programs on 14 conventional computers connected by a network there, using a Linda system built by Jerrold Leichter of Yale. One of those programs ran much faster on the network than it did on a supercomputer, a very powerful and expensive but serial (nonparallel) computer. (Other Linda programs did not perform as well.) Many organizations have large numbers of underused or idle computers in their internal networks. Those machines could well be used to search large data bases quickly.

The elements of a data base are ordinarily inspected one by one, but if a machine could look at them in the aggregate—converting many separate cases into a single "body of experience"—it might have a rich new source of knowledge. A smart data base is an information refinery that helps users to make sense of a new datum by bringing to bear its experience with many similar cases.

A conventional data-base system containing descriptions of breast X rays, for example, could retrieve all records for patients of a given age who have a particular diagnosis or records for a particular patient. The system we are developing attempts to fill in blanks in data items. If a doctor has a complete description of the image of a mass on the film, the system might be able to guess whether it is benign or malignant based on patterns in similar cases. The program can respond (in the same vein) to vague inquiries of the form "Get me all cases that resemble this one."

Furthermore, our smart data base attempts to mimic (albeit crudely) the creativity of human experts. Real experts don't merely interpolate between known data points and stick

CASE INFORMATION	SYSTEM SPECULATION AND CONCLUSIONS	COMMENTS
(ID 30) (AGE 42) (MASS DENSITY HIGH) (MASS BORDER COMPLETE? NO)	Concluding (DUCTAL PROMINENCE? NO) (ARCHITECTURAL DISTORTION? NO)	
	Speculating: MASS DENSITY CHANGED? Guessing INCREASED—e.g. case ((id 14) (age 46) (diagnosis CA. INF. DUCTAL)) case ((id 20) (age 60) (diagnosis CA. INF. DUCTAL)) case ((id 50) (age 70) (diagnosis CA. INF. DUCTAL))	Although it has no information on changes in mass density for this case, the system quickly finds several similar cases, all cancerous, in which the mass density—the apparent density of the lesion as seen in the X-ray image—had increased over time.
	Speculating: MASS COMET? Guessing YES—e.g. case 14 case ((id 40) (age 69) (diagnosis CA. INF. DUCTAL))	Similarly, the system speculates from similar cases that there is a mass comet (a particular type of image around the lesion). These speculations imply malignancy.
(MASS TYPE BORDER LOBULATED) (MASS BORDER DEFINED? YES)	Concluding (LARGE CALCIFICATIONS? NO) (NIPPLE INVERSION? NO)	From the characteristics of the border of the mass seen on the X ray, the system deduces other features; for example, concluding that there is no nipple inversion means that if the mass is malignant it is probably either not very far advanced or not close to the breast surface.
(MASS DENSITY CHANGED? NO) (MASS HALO? NO)	NO is unusual in this context—see e.g. case 14 case 40	The combination of these two characteristics is rare; the first suggests a benign growth, whereas the second suggests a malignancy. The two cases cited are the only other films that lacked a halo and yet displayed the same constellation of other features.
(BACKGROUND DENSITY FATTY) (FAMILY HISTORY CANCER NO) (PERSONAL HISTORY CANCER MYELOMA)	!! Never seen this value before	Myeloma is a very rare lesion in the breast, so it is not surprising the system has not seen it before.
	Closest cases in data base: ID — Malignant? — Diagnosis 4 — NO — FIBROADENOMA 6 — YES — CA. COLLOID 40 — YES — CA. INF. DUCTAL 16 — YES — MET. MELANOMA 29 — NO — FIBROADENOMA 12 — NO — FIBROADENOMA	The system considers the other cases that show similarities to this one. Three of the most similar cases are malignant (colloid cancer, infiltrating ductal carcinoma and metastatic melanoma); the rest are benign.
	Speculating: DIAGNOSIS Guessing FIBROADENOMA Checking expectations PERSONAL HISTORY CANCER: NO expected MYELOMA found instead. Withdrawing this guess.	Based on diagnoses in cases with similar characteristics, the system speculates that the current case is a particular kind of benign lesion. But in cases of benign lesions the system expects to see no prior personal history of cancer. Finding such a history, it withdraws the tentative diagnosis.
	Speculating: MALIGNANT? Guessing YES Checking expectations MASS HALO?: NO expected Confirmed. Accepting this guess.	Although the system cannot determine the specific kind of this growth (it has never seen a plasmacytoma before), it can still make a judgment as to whether it is malignant. Comparison with similar cases and deductions from those comparisons implies malignancy. The tentative diagnosis is confirmed by the absence of a mass halo (which would indicate a benign tumor) around the growth.

"SMART" DATA BASE reasons about a possible case of breast cancer on the basis of similarities between a description of the X-ray film in this case and descriptions of cases whose diagnosis is known. The existing system is still in prototype form and has access to only about 70 cases. Here it is confronted with a malignant tumor (plasmacytoma) that it has never encountered; it is unable to come up with a precise diagnosis, but it does recognize that the growth described is malignant.

to what is statistically solid; they make guesses, form hypotheses and jump to conclusions. Our system is designed to attempt something similar. It should alert users to interesting possibilities even when they cannot be firmly established (and might very well be wrong). We call this kind of behavior "simulated speculation." (The anthropomorphic term is used only for conciseness; the program does not attempt to mimic human psychological processes.)

The system is intended to grow every time it is used. Whenever the user presents a query to the data base, the query can be saved as if it were another case. When new data or new queries come in, the system can check whether they shed light on old ones. The system might say, in effect, "The new case you've entered recalls a query from XYZ six months ago and may offer a new perspective."

The prototype program (currently designed for sequential rather than parallel execution) is called the FGP machine, for fetch, generate and project. It was written by Scott Fertig of Yale, based on earlier work by Mitchell Sklar, with contributions from Paul Fisher. Items in the FGP machine's data base are represented as regions in space. Nearby regions correspond to items that are similar to one another. (This model draws on well-established approaches to text retrieval.) When presented with an inquiry, the machine creates a region corresponding to the characteristics of the query and then examines its neighborhood for closely related cases. The machine then inspects those cases for shared attributes—those likely to hold true for the region representing the query as well. These attributes, along with their comparative likelihoods, can then be presented to the user.

Having reached whatever conclusions seem reasonable, the machine can now engage in a bit of simulated speculation. Turning aside from the inquiry at hand, it focuses on any "evocative" possibilities that may have suggested themselves during the examination of nearby regions—unknown facts that could substantially narrow the focus of inquiry if they were known. For example, the initial presentation of a breast X ray might suggest the possibility of a mass comet—the appearance of a cometlike trail around a growth. Mass comets are useful diagnostic indicators, and so the system would shift its focus and consider where and when they are generally found. The machine moves to the region of the data base represented by the evocative possibilities and examines the new neighborhood just as it examined the neighborhood of the original query.

The degree to which the FGP machine allows itself to be distracted, so to speak, depends on a parameter called the concentration level. When the concentration is set at maximum, the program never speculates; at lower levels, it wanders. When the concentration level is set at zero, it free-associates, moving from one interesting possibility to another without ever returning to the original query (except by chance). Most people, of course, would have no use for a machine that ignored their questions and spent its time free-associating; this capability is merely a curious (and perhaps thought-provoking) side effect of the machine's design.

Unlike the ICU watcher or the mass-search machine, the FGP machine is not now a parallel program. It runs so slowly, even when working with small data bases, that it is almost unusable. We have already begun rebuilding the FGP machine in parallel form. The advent of relatively cheap parallel hardware and—more important—straightforward and effective techniques for building parallel software machines has made it possible to continue a project that would have been (in practical terms) a dead end 10 years ago.

We will use techniques like those of the mass-search machine to speed up the FGP machine's execution so that the machine can calculate the distances between the query region and many regions of the data base simultaneously. Then we will call on network-programming methods related to those demonstrated at Sandia to allow FGP machines to draw on data bases all over the country (or the world). When the technology has been fully developed, it will be possible to extract useful information from data bases that hold descriptions of various objects or events—be they wildflowers, cuneiform tablets, criminal records, news stories or vacation retreats. A user will be able to ask: "We are trying to identify a criminal (or flower) with the following characteristics. Who (or what) might match them? If there are no good matches, what other characteristics might such a person (or flower) have?" Ultimately we intend to focus the FGP machine on diverse and heterogeneous collections of facts; it will serve as a sophisticated interface to an encyclopedia, then perhaps to a library, then perhaps to all the libraries in the country.

A great deal of work remains to be done before such plans are realized. The possibilities, however, are virtually endless. Beyond the transportation watcher and the intensive-care monitor lies a different way of viewing large systems: the software microcosm project. Researchers in our group are attempting to build software "shadow systems" that track the evolution of actual systems or organizations (such as hospitals or universities), perhaps by building software machines containing a module to track the evolution of each separate unit of the system being modeled. The software microcosm itself would do a minimum of specific analysis, but the user would be able to ask questions, monitor the quality of work being done, file standing requests for data or just poke around in the corners looking for information.

The goal of all these projects—the trellises, the smart data bases, the microcosm—is essentially the same: to put a vast world inside a small box so that people can explore it any way they choose. Techniques developed by our research group and by many others are showing in concrete terms how to proceed. Before long, though, projects like these will outstrip the bounds of any academic research effort. If they are to proceed, other organizations with more resources will have to carry the work forward.

Unfortunately, government, industry and even the scientific community have only a vague feel for software. The word "metamorphosis" was put in the title advisedly; the word itself was metamorphosed by Kafka's creature. The new technology will lend itself to abuse, to charlatanry and to Luddite attacks. For their own protection, edification and delight, those who do not build software must come to grips with the potential of the new information machines.

FURTHER READING

DISTRIBUTED INTERPRETATION: A MODEL AND EXPERIMENT. Victor R. Lesser and Lee D. Erman in *IEEE Transactions on Computers*, Vol. C-29, No. 12, pages 1144–1163; December, 1980.

CONNECTIONIST EXPERT SYSTEMS. Stephen I. Gallant in *Communications of the ACM*, Vol. 31, No. 2, pages 152–169; February, 1988.

GETTING THE JOB DONE. David Gelernter in *Byte*, Vol. 13, No. 12, pages 301–308; November, 1988.

LINDA IN CONTEXT. Nicholas Carriero and David Gelernter in *Communications of the ACM*, Vol. 32, No. 4, pages 444–458; April, 1989.

The Middle-Ear Muscles

Tiny muscles behind the eardrum contract involuntarily when a person vocalizes or is exposed to a loud noise. This neuromuscular control system prevents sensory overload and enhances sound discrimination

by Erik Borg and S. Allen Counter

Modern industry has produced a noisy world. The din of jackhammers, the whine of jet engines and the blare of amplified electric guitars have become all too commonplace. It was therefore considerate of nature to have equipped the human ear with a rather sophisticated noise-reduction system: two small muscles that are attached to the ossicles, the tiny bones that connect the eardrum to the cochlea (the structure that houses the sound-receptor cells). When the muscles contract, they dampen the vibrations of the ossicles, thereby reducing the acoustic signal that ultimately reaches the inner ear.

Although they are skeletal muscles (in fact they are the smallest skeletal muscles in the human body), the middle-ear muscles are not under voluntary control. They contract reflexively about a tenth of a second after one or both ears are exposed to loud external sounds. Indeed, the characteristics of the reflex have become so well known that deviations from the normal response serve as a basis for diagnosing various hearing disorders and neurological conditions.

The muscles of the middle ear contract not only in response to loud external sounds but also immediately before a person vocalizes. This prevocalization reflex operates even when one speaks, sings or cries as softly as possible. Yet most evidence suggests that it is meant to protect the inner ear from the fatigue, interference and potential injury caused by one's own louder utterances, which can result in high sound levels in one's head. The shouting and wailing of children or babies, for example, can reach their own ears with the same intensity as the sound of a train passing nearby.

The middle-ear muscles do more than just indiscriminately attenuate internal or loud external sounds in humans. The muscles muffle primarily a loud sound's lower frequencies, which tend to overpower its higher frequencies. The net result of this frequency selectivity is to improve hearing—particularly of those sounds that contain many high-frequency components, such as human speech. In fact, the middle-ear muscles are what enables one to hear other people talking even while one is speaking.

Perceived sounds—regardless of their source—are air-pressure waves that have been funneled to the tympanic membrane, or eardrum, causing it to vibrate. The vibrations are transmitted through the three ossicles in the middle ear (the malleus, incus and stapes) to the cochlea. The middle-ear mechanism—the eardrum and ossicle linkage—serves to convert the movements of low-density air into analogous movements of the higher-density fluid in the cochlea. The movements of the fluid are transmitted to the stereocilia: fine, hairlike protrusions of receptor cells on the cochlea's basilar membrane. Mechanical forces on the stereocilia cause the cells to trigger electrical impulses in the auditory nerve that are then interpreted by the brain as sound.

Attached to the ossicles are the two middle-ear muscles: the tensor tympani and the stapedius [*see illustration on opposite page*]. The tensor tympani is connected to the neck of the malleus and is anchored in the wall of the eustachian tube (a ventilating tube connecting the throat, nasopharynx and middle ear). The stapedius originates in the wall of the middle-ear cavity and ends at the neck of the stapes, near its articulation point with the incus. The basic anatomy of the middle-ear muscles was described as early as 1562, by Bartolomaeus Eustachius (for whom the eustachian tube is named). Yet the function of the muscles in human hearing was a subject of speculation until this century, when laboratory experiments on animals and clinical observation made a comparative analysis of their physiology possible.

The middle-ear-muscle system is found in all classes of vertebrates, but it has distinctive features in certain species. In some species of frogs, for example, the hearing organ contains only a single ossicle that has a stapediuslike muscle attached to it. It is interesting to note that those frog species without a muscle or an ossicle in the middle ear tend not to vocalize.

Among lower vertebrates, birds possess the most elaborate systems for hearing and sound communication. In each ear they have a stapedius analogue, which is attached to both the tympanic membrane and a single ossicle, the columella. Because a bird's stapedius muscle lies mainly outside the middle-ear cavity, it can be studied more readily than the stapedius of mammals without damaging the delicate middle-ear structures.

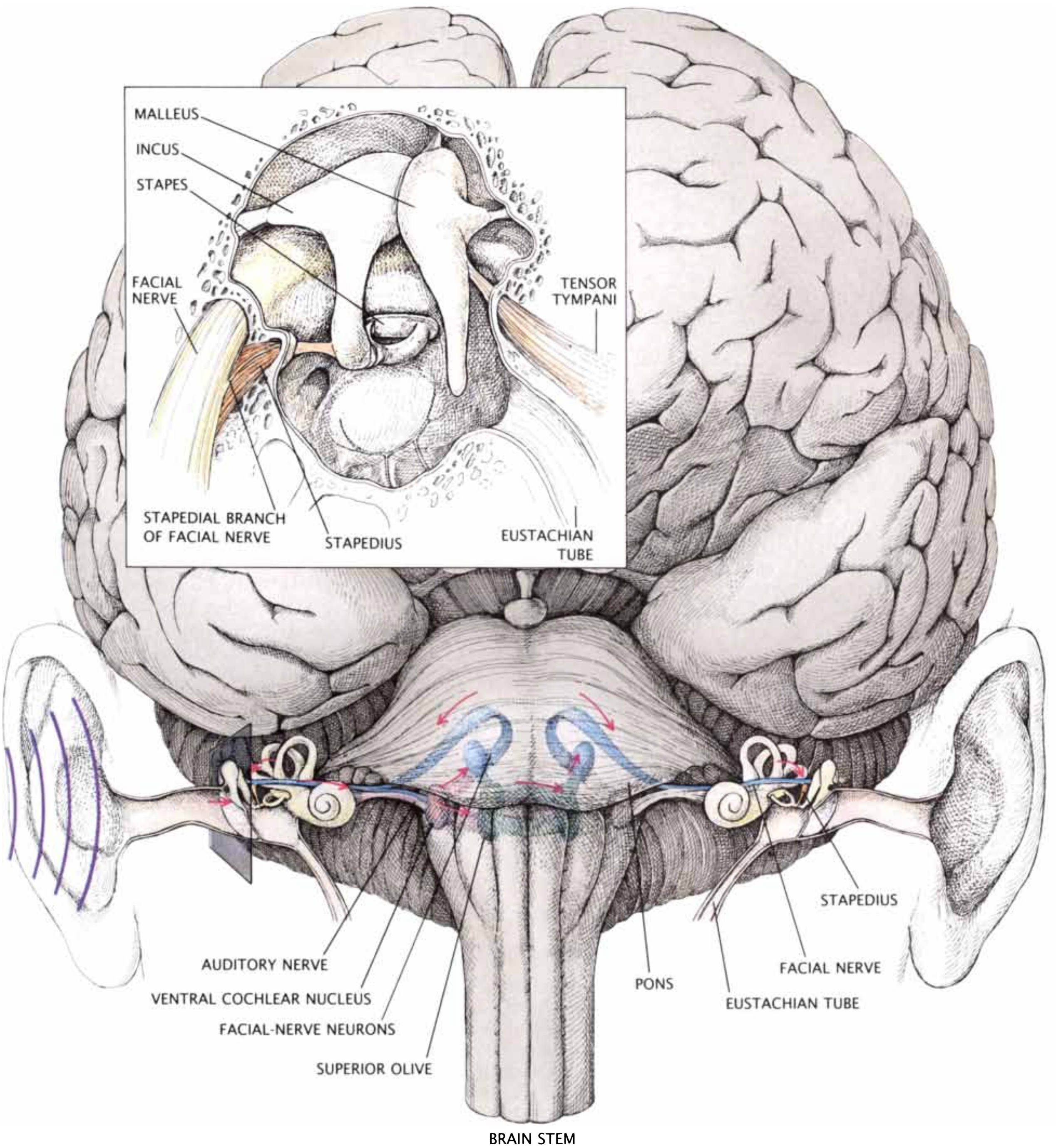
We have worked with common domestic fowl, such as chickens, in a series of experiments on the physiology of the stapedius at the Karolinska Institute in Stockholm and at Harvard University. By attaching a strain gauge to the tendon of a bird's stapedius and then stimulating the muscle electrical-

ERIK BORG and S. ALLEN COUNTER are long-standing collaborators in the study of the middle-ear muscles and brainstem physiology. Borg is head of the department of audiology and otology at the Karolinska Hospital in Stockholm and directs a laboratory at the Karolinska Institute for the investigation of experimental and clinical aspects of hearing disorders. He received his M.D. from the Karolinska Medical School in 1973. Counter is a neurophysiologist at the Harvard Medical School and the Massachusetts General Hospital. He received his doctorate from Case Western Reserve University in 1970. He is currently investigating the use of brainstem-evoked potentials and the acoustic stapedius reflex in diagnosing neurological disorders. Counter is also director of the Harvard Foundation.

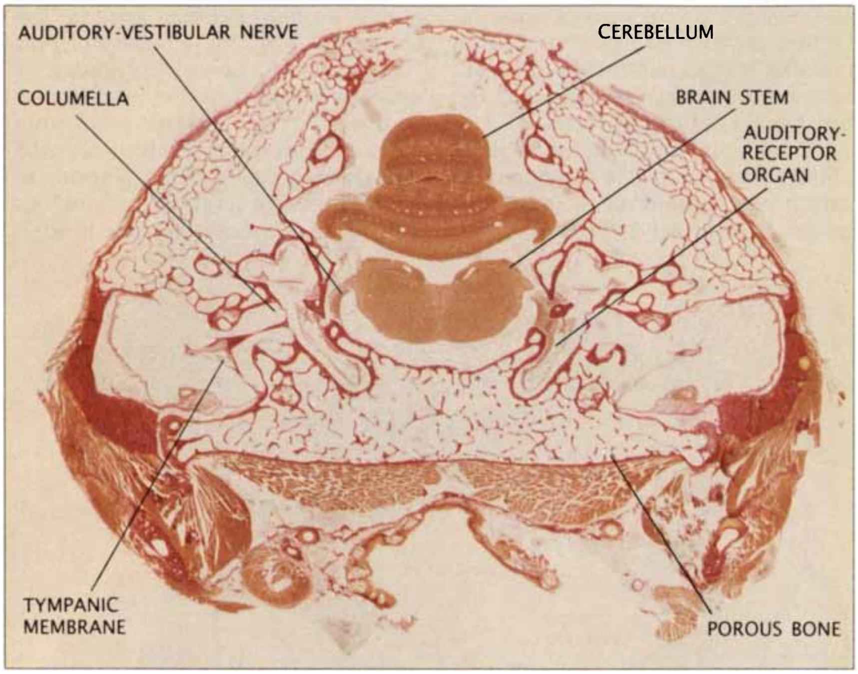
ly, we found that the stapedius was capable of contracting at rates in excess of 100 times a second.

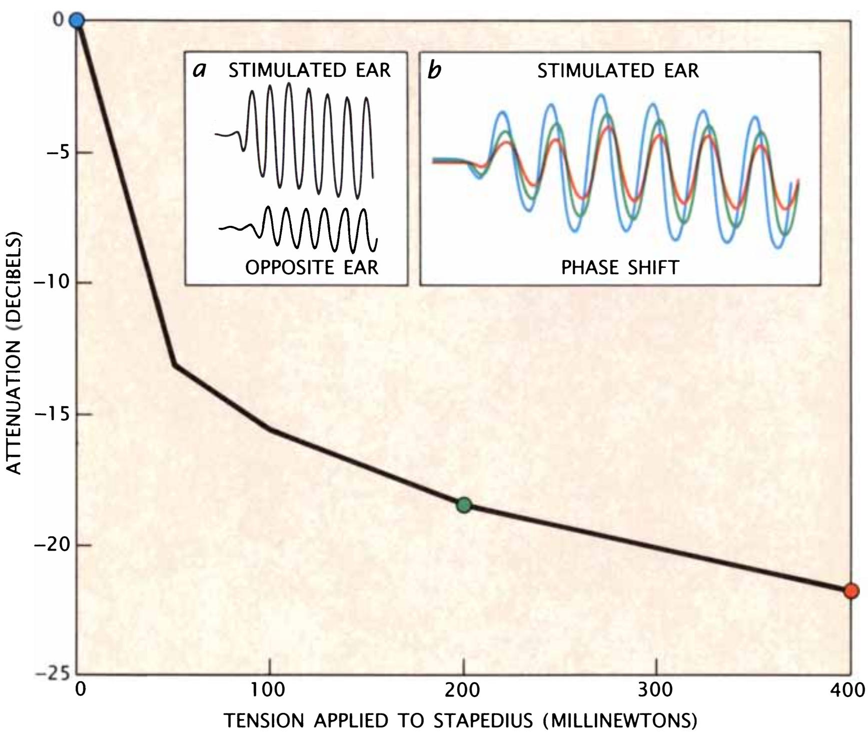
The muscle's inherent capacity for quick response and fatigue resistance is also evident from the microscopic appearance of its fibers. Electron micrographs show that the fibers contain abundant mitochondria (which provide energy), dense sarcoplasmic reticulum (which releases the calcium ions that trigger contraction) and numerous transverse tubules for the transmission of calcium ions.

Although the stapedius muscle of a bird always contracts during vocalization and swallowing, it does not appear to contract reflexively in response to loud external sounds. The primary role of the muscle in birds, then, seems to be the prevention of sensory overload of the auditory receptors during the birds' own loud cries. Indeed, the screech of seagulls and the crow of cocks can result in sound-pressure levels of as much as 130 decibels (measured at the head)—



MIDDLE-EAR MUSCLES (the stapedius and the tensor tympani) are seen in a view down an ear canal from which the eardrum has been removed (*inset*). When the muscles contract, they dampen the vibrations of the ossicles (the malleus, incus and stapes): the tiny bones that connect the eardrum to the cochlea, which houses the auditory receptors. The net result is a reduction of the sound transmitted to the cochlea. A loud sound (*purple*) in one ear activates the stapedius muscles of both ears through a neuronal pathway (*blue, red and green*) that lies in the lower brain stem (*above*). The pathway includes the auditory nerve, the ventral cochlear nucleus, the trapezoid body (not shown), the superior olive and the facial nerve.





STRUCTURE OF THE AVIAN HEARING SYSTEM is laid bare in a frontal cross section of a bird cranium (*top*). Birds have a single middle-ear muscle, the stapedius (not shown), and a single ossicle, the columella. Birds also have a porous-bone channel that connects both middle-ear cavities, which allows sound literally to go in one ear and come out the other. As a result the eardrum in the far ear normally vibrates out of phase with the eardrum in the ear directly exposed to sound. The phase difference is clearly evident (*bottom*) in a side-by-side comparison of the two ears' cochlear microphonics (*a*): the electrical output of their receptor cells, which mimics their acoustic input (in this case, a pure tone of 800 hertz). Experiments have shown that as tension is applied to the stapedius the sound reaching the inner-ear receptors decreases in intensity (*black curve*). Because the stapedius contracts vigorously while a bird vocalizes, the muscle's main function is probably to attenuate the sound produced at the bird's inner ear by its own cries. Also, tensing of the stapedius shifts the phase of an incoming sound wave slightly, as shown by cochlear microphonics (*b*). The attenuation and phase shifting might help birds to localize sound sources.

about the level of noise produced by a jet engine 15 meters away.

Direct evidence for the self-protection hypothesis came from recordings of a bird's cochlear microphonics: the electrical output of the cochlear receptor cells, which parallels the acoustic input to the cells. When we applied small amounts of tension to the stapedius muscle as we exposed the ear to a test tone, we observed a significant reduction in the amplitude of the cochlear microphonics and hence of the amount of sound energy reaching the receptors of the inner ear. The sound attenuation caused by stapedius-muscle contractions was essentially equal at all frequencies in the bird's auditory spectrum.

Another interesting observation was that tension in the avian stapedius muscle shifts the cochlear microphonics phase—the pattern of peaks and troughs making up its waveform—slightly with respect to time. What purpose might such a phase shift have? The answer has to do with the air-filled intracranial passageway that connects ears in birds and allows sounds literally to go in one ear and out the other. A sound wave crossing from one side of the head acts on the other side's tympanic membrane from the inside, generating cochlear microphonics in that ear that are slightly attenuated and almost totally out of phase with the microphonics of the ear through which the sound wave entered. We hypothesize that the contractions of a bird's stapedius muscles might modulate the binaural amplitude and phase relations in a way that helps the bird to locate sound sources.

One of the most specialized middle-ear-muscle systems is seen in echolocating bats. Their powerful stapedius and tensor tympani muscles contract at very high repetition rates as the bats make the rapid-fire click sounds that are their hunting cry. During each click a bat's middle-ear muscles reach peak tension quickly, but they relax just as suddenly so that the bat's ears will be sensitive to the click's echo from potential prey. The contraction-relaxation cycle lasts for only a few milliseconds (thousandths of a second) and can be repeated more than 100 times a second as the bat closes in on its prey. O'Dell Williams Henson, Jr., now of the University of North Carolina at Chapel Hill and Nobuo Suga of Washington University in St. Louis have confirmed that the bat's middle-ear muscles contract reflexively several milliseconds before the start of vocalization and, when contracted, can dampen the sound energy reaching the inner ear by more than 20 decibels.

Experiments on individuals who have perforated eardrums show that the middle-ear muscles are active during vocalization in human beings as well. By inserting a harmless needle electrode through the perforation and into the stapedius, one can make an electromyogram (EMG) of the muscle as the subject vocalizes. An EMG records the electrical activity of muscle fibers, which increases as the fibers contract. Such EMG's have revealed that the electrical activity of the stapedius begins just before the subject makes a vocal sound; the activity increases as the vocalization becomes louder. Presumably, then, the stapedius contracts automatically in the ears of both whispering mothers and crying babies—from the softest to the loudest vocalized sounds. The tensor tympani probably contracts before vocalization as well.

The middle-ear muscles can also become active spontaneously, in the absence of any sound, resulting in the perception of a repetitive and often bothersome noise. In addition, tactile or electrical stimulation of certain skin zones on the face and ear can elicit a contraction in the stapedius.

In contrast to birds, human beings and mammals in general appear to have another type of reflexive contraction of the middle-ear muscles—one evoked by loud external sounds. In most mammals both the tensor tympani and the stapedius muscles are subject to such a reflex, but in humans only the stapedius is. For that reason the middle-ear response to loud sounds in humans is called the acoustic stapedius reflex (ASR).

The ASR causes the stapes to move some 50 microns (millionths of a meter) from its resting position, thereby increasing the stiffness of the ligaments holding the ossicle, which in turn reduces sound transmission to the inner ear by 20 decibels or more. Like the pupillary reflex (the contraction of both pupils in response to light shone in one eye), the ASR is normally observed in both ears, even if only one is stimulated acoustically. The reflex is generally elicited by sounds that are between 80 and 90 decibels above a person's hearing threshold—about as loud as the sound of a noisy street.

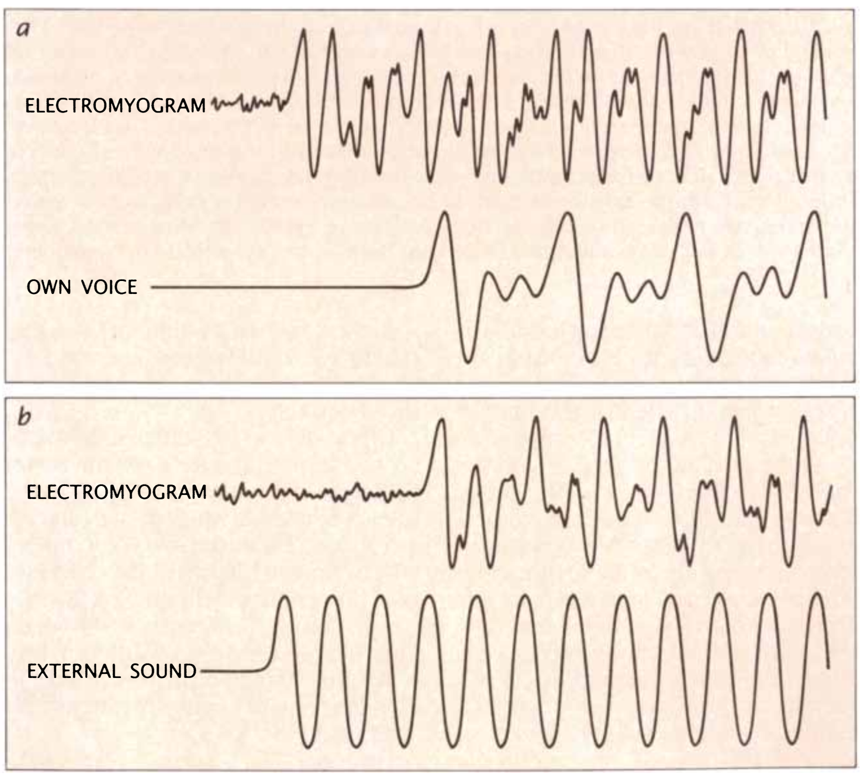
The importance of the middle-ear muscles in human hearing has been a matter of considerable debate over the years. Several single-function theories have been advanced, but the findings of many studies suggest that nature has been economical: it has given the muscles several separate but interrelated functions.

Animal-model studies of the middle-ear muscles conducted in the early 20th century by the Japanese otolaryngologist Toru Kato, along with more recent studies of the ASR by us and others, have made it fairly clear that the human stapedius is capable of protecting the inner-ear receptor cells from sustained, loud noise that might otherwise cause hearing loss—particularly in the frequency range that is most important for speech communication. Such hearing loss occurs when the inner ear suffers an acoustic battering that fractures stereocilia and thereby incapacitates receptor cells.

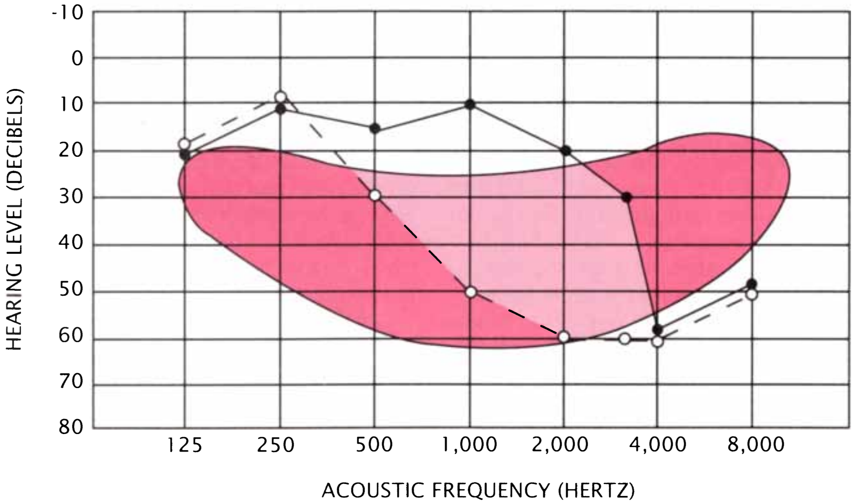
The stapedius cannot protect the inner ear from the damage that can be caused by an exceedingly sharp and intense sound pulse, however. The stapedius requires between 100 and 200 milliseconds to contract fully—a response time that is too slow to muffle, say, the sound of a gunshot before it reaches the inner-ear receptors. (Actually, one can probably reduce the risk of inner-ear damage from the bang of a gunshot by humming before shooting, since the middle-ear muscles are activated automatically during vocalization.) The stapedius can attenuate loud, abrupt sounds only if they come in quick succession, since the muscle then has a chance to build up tension. Apparently, the middle-ear muscles evolved to cope with the sounds of nature, such as thunder and loud animal sounds, which tend to rise slowly. Nevertheless, they do a remarkable job of protecting one's hearing from much of the noise of modern industrial society.

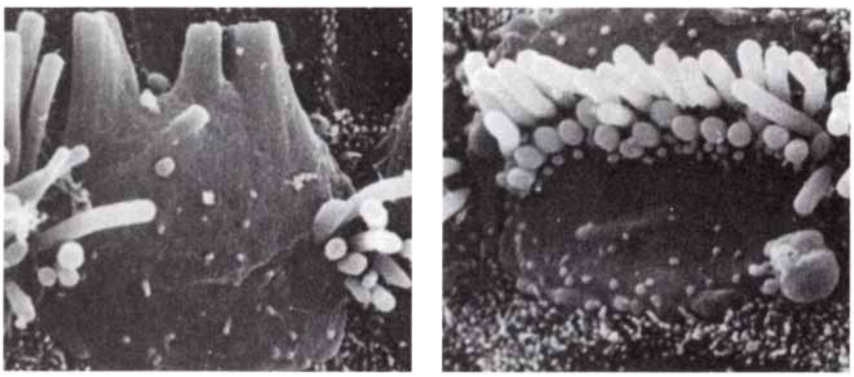
Other recent investigations have shown that the stapedius muscle has an even more sophisticated role in human communication. Audiologists have noted that people with nonfunctional stapedius muscles tended to have some difficulty in distinguishing speech sounds when loud background noise was present or when the sounds were amplified greatly. The lack of functional stapedius muscles appeared to have hampered the subjects' ability to discriminate spoken words. How could that be?

The answer becomes clear when one considers the way a sound wave is broken down into component frequencies in the long, spiral cavity of the cochlea. A sound wave propagating in the inner-ear fluid generates a traveling wave along the basilar membrane, which partitions the cochlear duct from beginning to end. Low-fre-



ELECTROMYOGRAM (a tracing of the electrical activity of muscle fibers) that is recorded in step with tracings of internal and external auditory stimuli shows that the human stapedius contracts involuntarily just before a subject vocalizes (*a*) and just after the subject is exposed to a loud sound (*b*). (The tracings are schematized.)





AUDIOGRAM (*top*) shows the clinical consequence of an inactive stapedius. The hearing of a person with a paralyzed stapedius who has had prolonged exposure to high levels of noise (*broken line*) is significantly worse than the hearing of a person in a similar auditory environment but who has functioning middle-ear muscles (*solid line*), particularly at frequencies ranging from 250 to 4,000 hertz. That frequency range (*light red*) includes many frequencies that form a part of normal speech (*red*). The physical cause of such hearing loss is the fracture of stereocilia (*bottom left*): the microscopic hairlike sensors of the auditory receptor cells. Normally the stereocilia are arranged in tapered bundles (*bottom right*). The micrographs were made by Berit Engström and Agneta Viberg of the Karolinska Institute in Stockholm.

quency sound components induce undulations in the membrane from the stapes all the way to the tip of the cochlear spiral, whereas the undulations of high-frequency components die out much more quickly and are confined to the area near the stapes. Because the low-frequency components dominate the undulations of the basilar membrane in the cochlea, low-pitched sounds of high intensity can drown out high-frequency sounds and even make them imperceptible.

Such "masking" of high-frequency sounds by low-frequency sounds is minimized by the ASR, since increasing the stiffness of the middle-ear linkage attenuates the low-frequency components of a complex sound more than it does the high-frequency components. Because many key speech sounds are generally high-pitched, the middle-ear muscles can actually enhance the perception of speech when they contract.

The ability of the stapedius muscle to maintain the ear's sensitivity to the frequencies encompassed by most speech sounds in spite of high sound levels was demonstrated experimentally by Roland Nilsson of the University of Gothenburg and John-Erik Zakrisson of Umeä University in Sweden. They showed that the ASR can improve the threshold for the detection of high-frequency sound in noise by as much as 50 decibels.

The stapedius also enhances one's ability to hear while speaking. A speaker's own ears are subjected to intense low-frequency vibrations that arise primarily from the enunciation of vowels. Fortunately, the prevocalization contraction of the middle-ear muscles prevents one's own speech from masking ambient high-frequency sounds. Indeed, the muscles are what makes it possible to hear soft sounds while one speaks.

The neural circuits that control the ASR and the prevocalization reflex are only now beginning to be understood. Modern techniques for the visualization of neurons in combination with the physiological studies have helped chart the intricate neuronal network that activates the middle-ear muscles.

The ASR relies on a complex pathway through several brain-stem nuclei (clusters of nerve cells) as well as the auditory and facial-nerve neurons [*see illustration on page 75*]. The primary neuronal pathway that controls the ASR in the ear being stimulated originates at the cochlear receptors, extends along the auditory nerve to the brain stem, where it includes parts of the ventral cochlear nucleus and the superior olive, and follows the facial nerve to its stapedial branch. There is evidence that the ASR is activated in the opposite ear by neurons from the ventral cochlear nucleus that communicate with the superior olive on the other side of the brain.

The close relation between the auditory neuronal pathway (which processes the signals from the receptor cells in the cochlea) and the motor neurons controlling the stapedius is particularly evident in experiments in which the stapedius is injected with labeling agents such as stains or certain viruses. These agents trace the neurons that make up the ASR pathway from the muscle to the brain stem. Such experiments have demonstrated that the cell bodies of the neurons innervating the stapedius in mammals (which exhibit the ASR) lie at the margin of the facial-nerve nucleus, very near the superior olive and the ascending auditory tract, from which they are activated. In contrast, birds (which exhibit no ASR) have stapedius motor neurons whose cell bodies lie within the facial-nerve nucleus, some distance from the ascending auditory pathway. It seems clear that these morphological differences in the brain stems of mammals and birds reflect the presence or absence of the ASR.

The neuronal pathway that controls the activation of the middle-ear muscles during self-vocalization has not been completely identified in mammals, but it probably has elements in common with the pathway that controls the muscles of the larynx, or voice box, during speech. Indeed, a reflex circuit from laryngeal receptors to the stapedius muscles has been found in some mammals.

It turns out that the neuronal pathway of the ASR travels through an area of the brain stem containing control centers for many vital physiological functions. Hence, the reflex offers a way to test the integrity of such brain-stem centers. A noninvasive diagnostic technique based on the ASR, called the acoustic-impedance-change test, can pinpoint the site of lesions on the cranial nerves or in the brain stem.

The method, first developed by the Danish physician Otto Metz in 1946, measures the ASR's effect on the vibration of the tympanic membrane and ossicles. The membrane is set vibrating by a continuous pure tone emitted from a small probe that is inserted into the ear canal. (The tone is not loud enough to elicit the ASR.) The probe, whose rubber tip ensures an air-tight seal, carries a miniature microphone that can measure the sound level reflected from the eardrum in the sealed ear canal. A second, loud tone (called the activator or the eliciting stimulus) is then generated at either the same or opposite ear in order to induce the ASR. When the stapedius contracts, the eardrum becomes stiffer and the reflected-sound level in the ear canal changes. That change (which represents a change in what is known as acoustic impedance) is recorded by the microphone, whose output is presented graphically as a tracing [*see illustration on next page*].

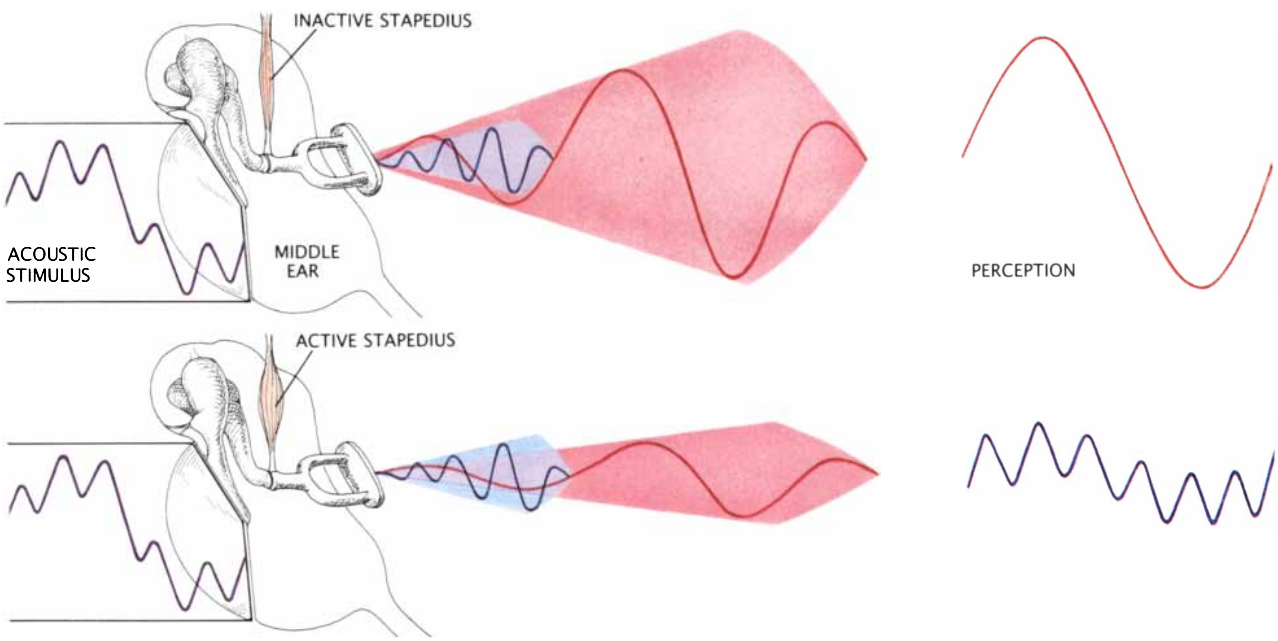
Normally the stapedius contraction produced in this fashion can be sustained for several seconds, during which the tracing shows little or no decrease in acoustic impedance. Henry Andersson of the Karolinska Hospital in Stockholm has shown that lesions on the auditory nerve can cause the ASR to decay at an abnormally rapid rate. Patients who have tumors near the nerve, for example, exhibit impedance tracings that decrease from the peak amplitude to prestimulus levels in a few seconds. The reason is that the tumor exerts pressure on the nerve fibers, which makes them more susceptible to fatigue.

Brain-stem lesions at the ventral cochlear nucleus are also manifested by abnormal decays in ASR amplitude and an increased ASR threshold in both ears. In contrast, lesions farther along the auditory pathway may block the reflex in the opposite ear but should not affect the response in the stimulated ear. Abnormal ASR magnitudes and decay rates are also exhibited by patients suffering from multiple sclerosis, a neurological disease.

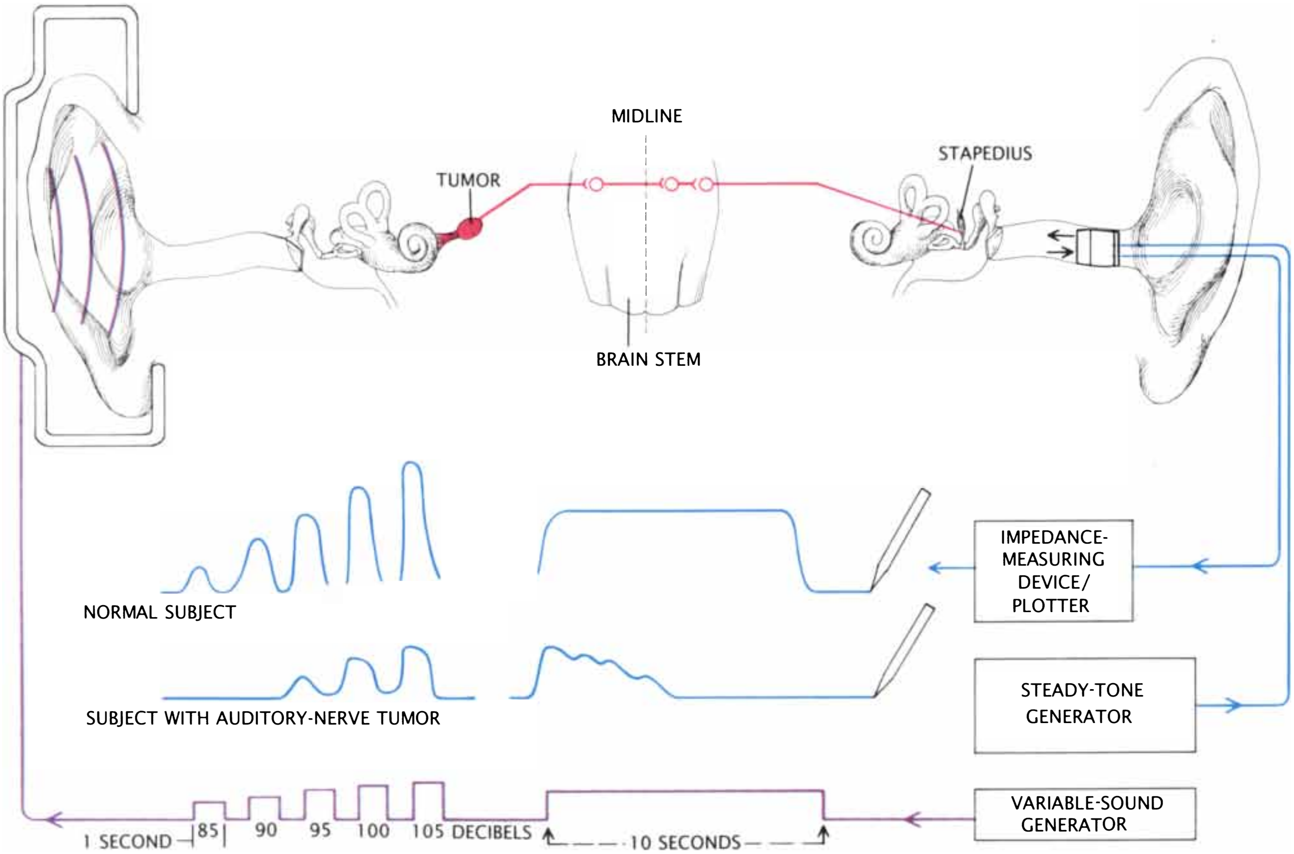
In patients afflicted with Bell's palsy (facial paralyis), the ASR is reduced or blocked entirely in one ear if the lesion lies between the facial nerve's origin and the branch innervating that ear's stapedius. In such a case, the acoustic-impedance measurements can serve to monitor the recovery of facial-nerve function during treatment of the condition or after spontaneous recovery. (Incidentally, patients who have a paralyzed stapedius muscle as a result of Bell's palsy have helped elucidate the muscle's role in human hearing: they complain that their ears are hypersensitive to loud sounds and that what they hear is often distorted.)

Measurements of acoustic impedance are also valuable in the diagnosis and monitoring of myasthenia gravis, an autoimmune disease that is characterized by muscular weakness and extreme susceptibility to fatigue. The disease is caused by the production of antibodies to the patient's own acetylcholine receptors on the muscle-surface membrane. Acetylcholine, a neurotransmitter, normally stimulates muscle activity; without functional acetylcholine receptors a muscle can quickly atrophy.

Acoustic-impedance recordings of patients with the disease typically show an increase in the ASR threshold, a reduction in its magnitude and an abnormally rapid decay. If more acetylcholine is made available by injecting the patient with a substance that inhibits the neurotransmitter's breakdown, the threshold drops consider-



ATTENUATION of a loud complex sound's low-frequency components by the stapedius helps to prevent the "masking" of its high-frequency components. Masking is thought to arise from the way a sound is broken down into high- and low-frequency components in the cochlea. In a person who has a nonfunctional stapedius muscle (*top*), the low-frequency components (*blue*) drown out the high-frequency ones (*red*), as is shown in a superposition of the two components' amplitude envelopes. Normally the stapedius minimizes such interference by dampening low-frequency sound components before they reach the inner ear (*bottom*), thereby allowing both high- and low-frequency sound components to be perceived.



ACOUSTIC-IMPEDANCE-CHANGE TEST takes advantage of the fact that the acoustic stapedius reflex (ASR)—the involuntary contraction of the stapedius muscle in response to a loud sound—can occur in both ears even when the sound is directed into only one ear. In the test, a headphone at one ear produces a sound loud enough to elicit the ASR in both ears. A probe then records the reflexive response in the opposite ear. The probe contains a small sound source that emits a steady, soft tone and a sensor that monitors the sound reflected from the eardrum. When the stapedius contracts, the ossicular chain and eardrum stiffen, and the sound reflected from the eardrum increases dramatically. In a typical run, tracings of the changes in the reflected-sound level (*blue*) are recorded as the ASR-eliciting sounds (*purple*) are progressively increased in intensity. The run is ended with a steady 10-second sound above the ASR threshold. The tracings for a healthy subject will mimic the pattern of ASR-eliciting sounds. In contrast, a person who has a tumor in the neuronal pathway on the side of the stimulus will exhibit weaker responses, an increased ASR threshold and a rapid decay in the response to the prolonged sound.

ably, and the ASR can be more than doubled in amplitude and duration.

The evolution of the middle-ear cavity and its associated structures endowed vertebrates with an increased sensitivity to sound. Yet that sensitivity, in turn, created a need for an efficient mechanism for coping with intense noises that would otherwise interfere with the perception of the sounds on which an animal's survival depends.

The middle-ear-muscle system is nature's elegant solution. This system of sound mufflers and tuners effectively suppresses loud internal and external noise, allowing relevant soft sounds to be separated from irrelevant loud ones. Specifically, the muscle's reflexive contractions prevent desensitization of the auditory receptors, interference between high and low frequencies in the perception of sound, and injury to the inner ear. It is quite likely that the significant evolutionary advantage conferred by the middle-ear muscles accounts for their existence throughout the phylogenetic scale, from the lower vertebrates to human beings.

The built-in reflexes of the middle-ear muscles have both enhanced the hearing of human beings and proved to be a reliable tool for determining the integrity of the ear and the neuronal circuits of the brain stem. Indeed, the acoustic-impedance-change technique has gained wide clinical application. It is likely that the application of the ASR will become even more important in the future for identifying individuals who are susceptible to noise-induced hearing loss as well as for rehabilitating those who have been outfitted with hearing aids.

FURTHER READING

ON THE NEURONAL ORGANIZATION OF THE ACOUSTIC MIDDLE EAR REFLEX: A PHYSIOLOGICAL AND ANATOMICAL STUDY. Erik Borg in *Brain Research*, Vol. 49, No. 1, pages 101–123; January 15, 1973.

PERIPHERAL CONTROL OF ACOUSTIC SIGNALS IN THE AUDITORY SYSTEM OF ECHOLOCATING BATS. Nobuo Suga and Philip H.-S. Jen in *Journal of Experimental Biology*, Vol. 62, No. 2, pages 277–311; April, 1975.

THE AVIAN STAPEDIUS MUSCLE: INFLUENCE ON AUDITORY SENSITIVITY AND SOUND TRANSMISSION. S. A. Counter and E. Borg in *Acta Oto-Laryngologica*, Vol. 94, No. 3–4, pages 267–274; 1982.

THEORIES OF THE MIDDLE-EAR MUSCLE FUNCTION. Erik Borg, S. Allen Counter and Günter Rösler in *The Acoustic Reflex: Basic Principles and Clinical Applications*. Edited by Shlomo Silman. Academic Press, 1984.

Maya Writing

The Maya had the most sophisticated script in pre-Columbian America. In the past decade scholars have finally been able to read it, filling in significant gaps in our knowledge of Maya society

by David Stuart and Stephen D. Houston

Prehistory is the study of peoples without writing, whereas history is the study of people who possess written texts. Yet, although the Maya had the most highly developed system of writing in pre-Columbian America, until recently they were studied largely as a prehistoric people. The reason was simple: their writing could not be read. Within the past decade all that has changed. Building on critical insights into the structure and content of Maya writing put forward in the 1950's, a fairly small group of scholars, including the two of us, has puzzled out a sizable fraction of the known writings of the Maya. These writings, inscribed on stone monuments or painted on ceramic vessels, have begun to augment the picture of Maya society that was derived from the patient work of excavation.

Some of the new findings confirm that rites of personal bloodletting, carried out by a noble elite, had an important role in Maya society. Less dramatic but perhaps more significant is what the inscriptions tell us about politics and geography. Most of the inscriptions are, after all, chronicles of specific rulers, marking their births, ascensions, rituals, conquests and deaths. Through careful study of these writings scholars are beginning to get a sense of the volatility of Maya politics, in which shifting alliances and wars among city-states led to rapid changes in the geopolitical landscape. This emerging picture is among the first fruits of a decipherment that, after a century of frustration, is bringing the Maya into history.

DAVID STUART and STEPHEN D. HOUSTON have been collaborating for some time in the study of Maya writing. Stuart has just graduated from Princeton University and moves to Vanderbilt University this fall to begin graduate studies. For the past several digging seasons he has worked on the inscriptions at Copán in Honduras. In 1984 he was awarded a five-year MacArthur Prize Fellowship for his work on Maya glyphs. Houston is assistant professor of anthropology at Vanderbilt. He received his Ph.D. in anthropology from Yale University in 1987; his dissertation focused on the politics of the Petexbatun region of Guatemala. Some of the problems raised in his dissertation will be explored in the next five years as he and his colleagues excavate four or five Classic Maya cities in the Petexbatun region.

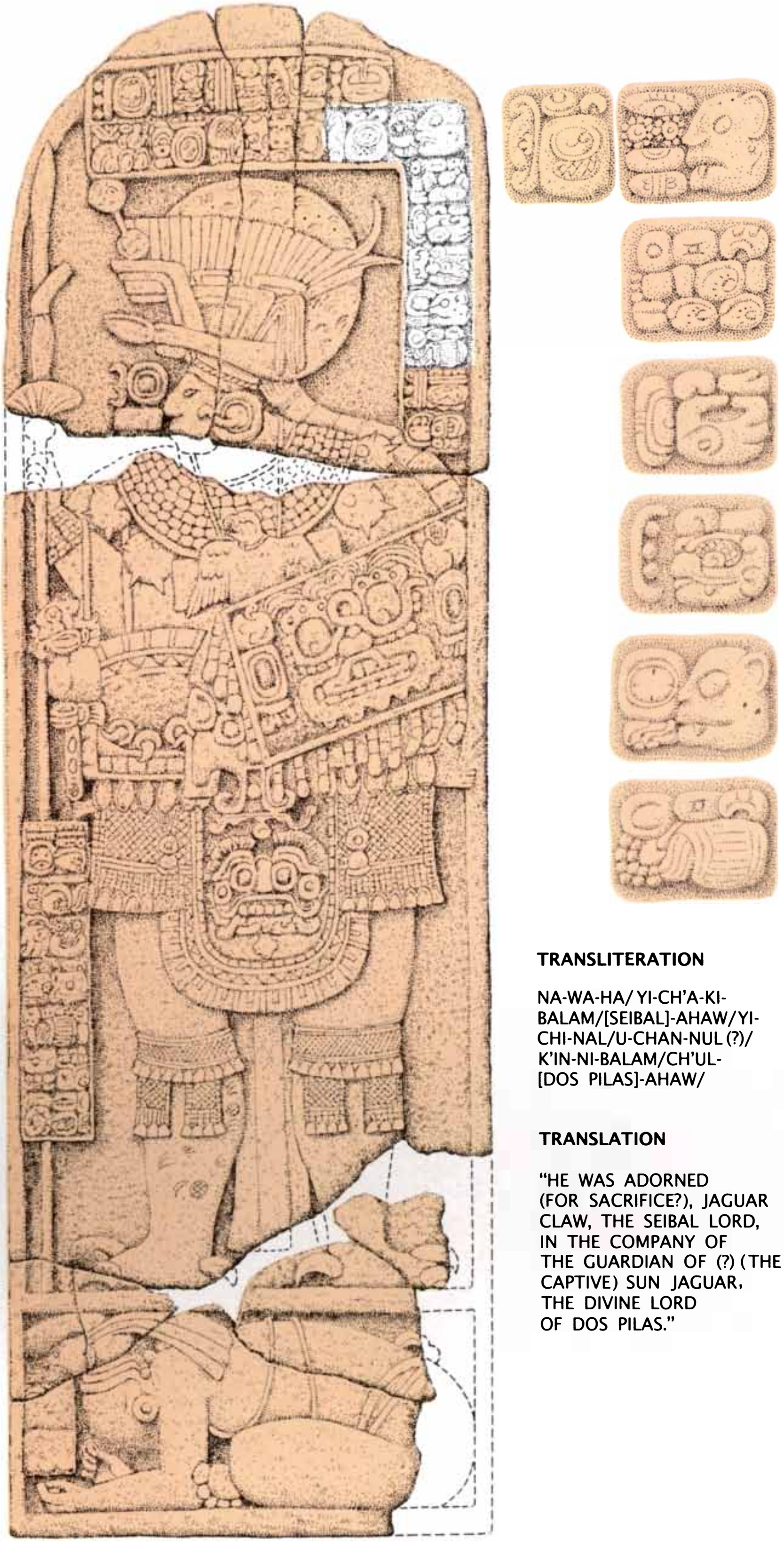
Set in the tropical region defined by what is now eastern Mexico, Guatemala, Belize and the western parts of Honduras and El Salvador, Maya civilization escaped thorough archaeological study until the beginning of the 20th century. Since then surveys and excavations have shown that by about 1500 B.C. people who probably spoke an ancestral form of Mayan had settled in the forested lowlands. For a thousand years the forest villages flourished largely unchanged. By about 250 B.C., however, fundamental changes were taking place in social life and political structure. The next two or three centuries witnessed the rise of powerful city-states whose lords, as the earliest hieroglyphic texts indicate, claimed for themselves a divine role on earth.

The glyphic system in which these early texts (and the subsequent ones) were written was not invented by the Maya. On the contrary, it seems they built on an ancient form that had been in use for centuries among advanced cultures to the west. Writing was present in what is now the state of Oaxaca by about 700 B.C., as is shown by the discovery there of a monument inscribed with early glyphs. A related system was developed in the modern state of Veracruz, as was demonstrated by the dramatic recent discovery of a stone monument inscribed with 420 glyphs at a site called La Mojarra. Although much of the text cannot be deciphered, two dates on it clearly correspond to A.D. 143 and 156. The writing on the La Mojarra stone may share antecedents with the earliest Maya hieroglyphs, dating from the second century A.D.

Whatever the precise origins of Maya writing, by the beginning of the Classic period of Maya culture in about A.D. 250 hieroglyphs were in use at hundreds of sites. During the Classic period, which lasted until A.D. 900 and represents the zenith of Maya culture, countless stone monuments had hieroglyphic signs carved on their surfaces. Writing also appeared in books made of bark paper and covered with jaguar skin, on painted and molded ceramics and on other portable objects. The Maya polities, in which such writings were produced, were generally small and possibly weak in infrastructure. Such infrastructural weakness may have contributed to the sudden collapse of Classic Maya society that occurred in about A.D. 900 (although the cause is still being debated by scholars).

In the Postclassic period that followed the collapse, writing was still used on some stone monuments and architecture and in bark-paper books, of which only four survive. Indeed, in the last unconquered Maya enclave (in northern Guatemala), writing probably continued until the 17th century. Today, although the written tradition has been lost, four million speakers of two dozen Maya languages still inhabit the Maya homeland. Their presence has

INSCRIBED STELA from Yaxchilan in Mexico dates from the seventh century A.D., at the height of the Classic period, which was the zenith of Maya civilization. The hieroglyphic text celebrates the capture of a foreign lord by the local ruler "Shield Jaguar." The carved scene depicts the lord and his kneeling captive.



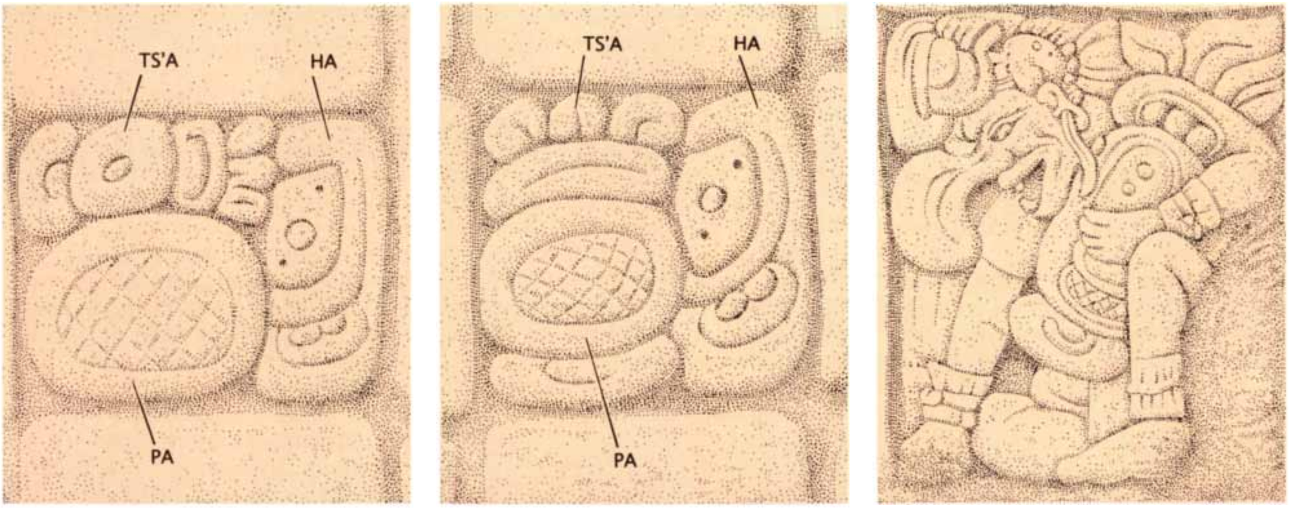
AGUATECA STELA 2 carries an inscription possibly indicating that one of the lords pictured—"Jaguar Claw" of Seibal—is dressed ceremonially for sacrifice. The standing figure is the ruler of Dos Pilas in Guatemala; the captive, Jaguar Claw, is shown beneath his feet. The political unit centered on Dos Pilas expanded rapidly in the late seventh and early eighth centuries A.D. to include Aguateca and other Maya sites.

been a boon for scholars studying Maya writing, because Colonial and modern dictionaries and studies by linguists have provided a way to check the decipherment as it proceeds, an advantage often denied those who seek to puzzle out ancient scripts.

The efforts at deciphering Maya writing actually began soon after the Spanish conquest of 1520. Indeed, the first students of Maya hieroglyphs were Spanish friars who were attempting to convert the Maya to Christianity. The most famous of these was Diego de Landa, third bishop of Yucatán, who in 1566 compiled a treatise called *Relación de las cosas de Yucatán* (roughly, "An Account of the Things of Yucatán"). In that work Landa included a brief summary of his understanding of Maya writing. In his view the glyphic signs were letters of an alphabet comparable to that of the Indo-European languages. He offered sketches of signs labeled A, B, C and so on. Unfortunately, what Landa had failed to perceive is that the Maya signs are anything but alphabetic. In any event, his treatise lay undiscovered by scholars for three centuries after it was written.

The first great scholarly effort at deciphering Maya writing was carried out by Ernst Förstemann, royal librarian at Dresden. In 1880 Förstemann began studying the hieroglyphs in the codices (the surviving bark-paper books, the most famous of which, the Dresden Codex, was close at hand) and the few inscribed stone monuments then known. Within 14 years Förstemann had unraveled the complex workings of the Maya calendar. He showed that the calendar is based on paired cycles of 260 and 365 days and that a date is generally expressed by noting its position in both the 260- and the 365-day cycle. Since that combination repeats itself every 52 years, the Maya chroniclers also recorded a date's position in a more precise "long count," a linear reckoning whose starting point is the year 3114 B.C.

Through the work of Förstemann and others, such as the American Joseph Goodman (who in 1905 proposed the correlation between the Maya and Christian calendars that is still widely accepted), it became clear quite early that the reckoning of time was of great importance to the Maya scribes. The notation of time, however, has many possible uses—including history, religion and myth—and it was not clear to scholars what was includ-



GRAPHIC VARIATION enabled Maya scribes to write each word in several ways. Shown here are three variants of the verb *ts'apah* ("was set upright"). Each example includes signs for three syllables: *ts'a*, *pa* and *ha*. In the first (*left*) the signs are in conventional order. In the second (*middle*) the *pa* sign has been inserted into the *ts'a* sign, which is vertical. In the third variant (*right*) the scribe has made use of a "full-figure" glyph for *pa*: a seated man with a bulbous nose who cradles a *ts'a* sign.

ed in the noncalendrical portions of the Maya inscriptions.

A first step toward understanding the noncalendrical inscriptions was made by another German scholar, Paul Schellhas, who identified many Maya deities and their names in the codices. Schellhas's work on the codices raised the possibility that the inscriptions might be largely religious or mythical in character, and this remained the view of many scholars well into the 1960's. As Maya archaeology expanded in the early decades of the 20th century, many new texts were discovered on stone and pottery, providing material for testing this hypothesis. Unfortunately, in spite of significant efforts by the leading epigraphers of the day—Sylvanus G. Morley and Sir J. Eric S. Thompson—little headway was made with the noncalendrical texts.

Curiously, none of the early attempts at decipherment made much use of Landa's intriguing "alphabet" after it was rediscovered in the 1860's. Perhaps the main reason for the lack of scholarly interest was that it seemed clear that at least some of the Maya glyphs were logographs: signs standing for entire words. For example, one of the 20-day "months" in the 365-day cycle is named for the bat, and the sign for the name of the month depicts a bat. Where Landa saw letters, early epigraphers saw whole words. As a result Landa's work languished.

Ironically, the first breakthrough in understanding the formal aspects of Maya writing was made by a young Soviet scholar who went back to Landa and took his work quite seriously. Yuri Knorozov of the U.S.S.R. Academy of Sciences in Leningrad accepted the conclusion that Landa's list was not an alphabet. But he did not reject it entirely for that reason. Instead Knorozov concluded that what had happened to Landa in making up his list was a form of intercultural miscommunication. For example, in asking his native informant to write the letter "B" (pronounced *beh* in Spanish) Landa had elicited the Maya sign for the syllable *beh*.

Knorozov proposed that Landa's list was actually a syllabary, or list of syllables. Each sign in the list stood for a specific combination of one consonant and one vowel. When the signs were joined, they phonetically spelled words; frequently the words were of the form consonant-vowel-consonant. Because few Maya words end in vowels, the final vowel would have been dropped in pronouncing the word. But when the word was written, the scribe would, according to Knorozov, have chosen a syllable that included the same vowel as the initial syllable. The principle of consonance between the initial and final vowels Knorozov called synharmony.

To test his theory, Knorozov turned to the codices. He began with a word that was thought on the basis of ancillary evidence to mean "turkey." In Yucatec Maya (a major group of modern Maya languages and one of those most closely related to the language of the ancient inscriptions), the word for turkey is *kutz*. Knorozov began with Landa's glyph for "K," which he interpreted as the syllable *ku*. The second sign in the pair of "turkey" glyphs was, on the principle of synharmony, likely to stand for the syllable *tzu*. He then turned to the two glyphs that made up a word in the codices that was thought to mean "dog." The first glyph was the hypothetical *tzu*. The second was Landa's "L," now interpreted syllabically as *lu*. *Tzul* (or *tzu-l(u)*, as the ancient Maya would have spelled it) was an old Yucatec word for dog.

Knorozov's work represented a remarkable breakthrough, and its fundamental principles—consonant-vowel (C-V) syllables and synharmony—are now accepted as valid. Yet for a variety of reasons it took many years after his work was done in the 1950's for it to be accepted in the West. In the meantime two Western scholars were making equally great strides in understanding the content of the inscriptions. Heinrich Berlin, an independent scholar who lived in Mexico City and supported himself as a businessman, pointed out that a certain category of glyphs seemed to stand either for places or for the ruling families affiliated with those places; these Berlin called Emblem Glyphs. As we shall describe, the Emblem Glyphs are now a significant focus of work on Maya writing.

In 1960, only two years after Berlin's work, Tatiana Proskouriakoff of the Carnegie Institution of Washington provided another turning point in the investigation of Maya writing. Proskouriakoff, who had entered Maya studies as an architect, was charting changes in Maya artistic styles. That work called for precise notation of the dates on monuments as a means of dating stylistic phases. The unexpected result was that the pattern of dates on the monuments corresponded to periods in the span of individual human lives. Using inscriptions from Piedras Negras in Guatemala as a case study, Proskouriakoff demonstrated convincingly that the recorded dates marked historical events in the lives of named rulers and their families.

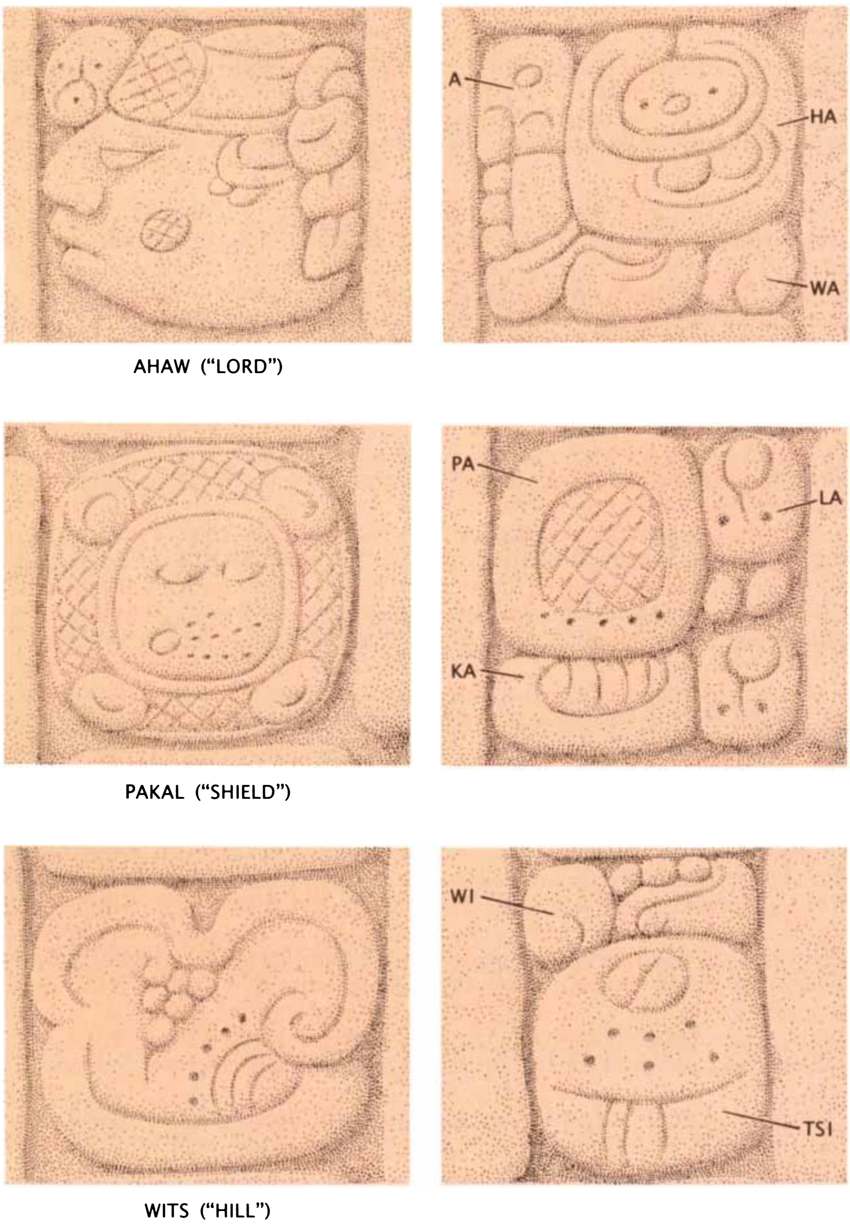
The contributions of Knorozov, Berlin and Proskouriakoff constituted a revolution that applied both to the form of Maya writing and to its content. For the first time it was understood that the writing system includes both logographs and signs for C-V syllables. At the same time it was understood that the content of the inscriptions generally applies to historical events in the lives of the ruling elite and not, by and large, to impersonal mythical or religious narratives. The implications of this revolution are still being worked out, and in the past decade progress has been particularly rapid.

As a result it is now possible to provide a tentative overview of the writing system and its contents. The basic elements of Maya writing are signs, of which about 800 are known. Individual signs usually have a square or an elongated oval appearance; one or more of them may be found together in what is known as a glyph block. Many such blocks are arranged in a rectilinear grid that provides the spatial framework for most of the known inscriptions. Within the grid, glyph blocks are arranged in rows and columns whose order of reading is prescribed by specific rules.

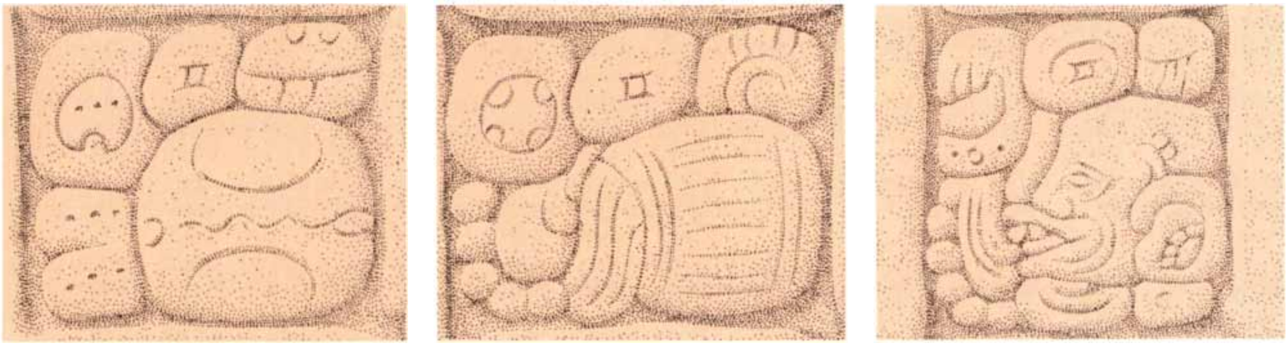
Signs are by nature highly pictorial, often representing in considerable detail animals, people, body parts and objects of daily life. The pictorial principle is taken to the extreme in inscriptions composed of "full-figure" glyphs, in which individual signs and numbers become animated and are shown interacting with one another. None of this should be taken to mean that the Maya had simple picture writing. On the contrary, the combination of C-V syllables and logographs enabled the scribes to write the words of their texts in detail.

Some of this flexibility comes from the availability of the two types of signs. For example, one very common honorific title in Maya texts is *ahaw*, meaning "lord" or "noble." *Ahaw* may be written in logographic form as a head in profile, with the distinctive headband or scarf that marked the highest nobility in Maya society. But it is also possible to write the word as a combination of three phonetic, syllabic signs: *a-ha-wa*. Likewise, the word *pakal* ("shield") can be indicated by a depiction of a shield or by the combination of syllabic elements *pa-ka-la*.

Because many Maya signs remain undeciphered, it is not possible to state precisely the relative proportions of logographic and syllabic signs. The number of deciphered syllabic signs keeps growing, and today about half of the syllabic grid is filled. (The syllabic grid plots the consonants of the spoken Maya language against its vowels and thus represents the totality of signs needed to write the language.) Half of the grid may seem a meager proportion, but it must be remembered that the discovery of the structure of the syllabic elements—Knorozov's main contribution—was made only a little more than 30 years ago. Furthermore, the C-V



LOGOGRAPHS AND SYLLABLE SIGNS offered the Maya scribes additional flexibility. A logograph is a sign for a whole word. Shown here are logographic and syllabic representations for the Maya words *ahaw* ("lord"), *pakal* ("shield") and *wits* ("hill").



EMBLEM GLYPHS are groups of signs associated with specific political units. They provided one of the first indications that Maya texts have historical themes, not mythical or religious ones. These Emblem Glyphs are from Dos Pilas, Palenque in Mexico, and Copán in Honduras. All three are based on the Maya title *k'ul ahaw,* meaning "divine lord." The first sign is *k'u* or *k'ul* ("holy" or "divine"); the second is *ahaw* ("lord"). The variable third sign presumably refers to the specific city-state.

syllables that are already understood are the common ones. Many of the empty spaces in the syllabic grid remain so because they are linguistically rare; rare signs are more difficult to translate than common ones.

Nonetheless, the pace of phonetic decipherment is bound to increase in the coming years as more resources are trained on it. One aspect of Maya writing that may complicate this progress is the fact that different signs can have the same value. Two signs that share a value are known as allographs. Such equivalences are common in Maya texts, and in evaluating a particular phonetic interpretation of a syllable, it is very helpful to identify as many as possible of the variant forms. The process of recognizing allographs depends on the slow work of comparing many texts in order to find variant spellings of the same word.

Allographs are not restricted to phonetic syllables. In a logographic form of substitution, more than one sign can be employed to represent the same word. For example, the Maya word *kan* (or *kaan*) may mean "snake," "sky" or "four," just as the English word "tie" may mean an article of clothing worn around the neck or an equal score in an athletic contest. The shared sound of these words provided the ancient scribes with a basis for word play. Indeed, the substitution of one sign for another sometimes approaches punning, as when a "sky" sign is found in a context where "four" is to be understood.

Such equivalences are not easily recognized, and they exemplify some of the difficulties that confront students of Maya texts. In spite of such problems, the pace of decipherment in the past few years has been particularly rapid. In fact, it has been so rapid that scholarly publication has sometimes failed to keep up with the proliferation of inscriptions newly translated in the field.

What has this recent progress revealed about the Maya themselves? In interpreting the information from the inscriptions, it must be kept in mind that the glyphs shed light on only one stratum of the pyramid of Maya society: the apex. Inscriptions were commissioned by the ruling elite, and they include only the information that the rulers thought was important. Such information is most welcome to scholars, but it must be kept in perspective. In relation to the great majority of Maya people—farmers, minor artisans, traders, masons—the written record is silent.

And what were the concerns of the ruling elite? Of paramount concern were lineage ties and political authority. Hints of this preoccupation became clear quite early. Berlin and Proskouriakoff uncovered the names of the rulers and their consorts at sites such as Palenque, Piedras Negras and Yaxchilan and constructed king lists. Later work has gone beyond lists of rulers to clarify the family relationships among the people named in the inscriptions, and it has become clear that during the Classic period rule in Maya society was passed from father to son, much as it was in the hereditary monarchies of Europe.

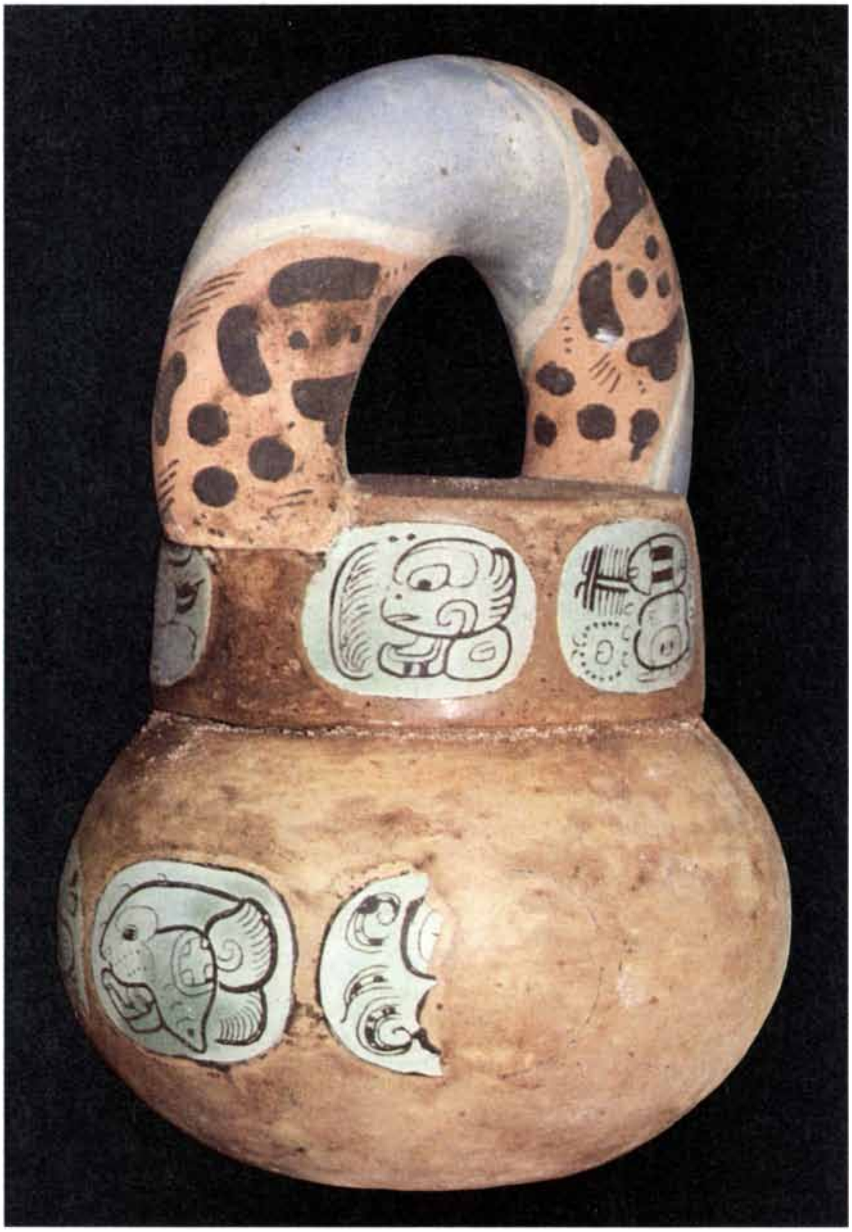
Much space in the inscriptions is devoted to relations between fathers and sons in the ruling lineages and to other kin relations, and it can be assumed that kinship was a topic of special importance to Maya rulers. Family connections appear to have been fundamental to the political organization of Maya society. Marriage between ruling lineages of different polities had an important role in diplomacy and in the forging of alliances. Within individual polities, members of the royal family who were not in direct line to the throne sometimes filled bureaucratic roles.

Other members of the nobility became skilled craftsmen, as is shown by a ceramic vessel from the Naranjo area of northern Guatemala. This vessel, which unfortunately was looted (a process that deprives indigenous peoples of their heritage and archaeologists of crucial information about origins of artifacts), is signed by its maker. The signature reads in part: "the son of the Naranjo *ahaw* and the Lady of Yaxha." Many artists left their signatures on pottery and on stone monuments. More than one name frequently appears on a single sculpture, which testifies to the collaborative effort required to make large works as well as to the value put on the work of celebrated artists.

As the preceding paragraphs suggest, many rulers, their relatives and their immediate subordinates can now be identified by name and by position in the rigid Maya social hierarchy. Yet such identifications provide only the bare outline of a culture. Of the rest—how those so identified actually spent their time—only tantalizing glimpses are available. Most Maya texts describe only major episodes in the lives of rulers and only those that bear directly on their status as lord, such as birth, accession to the throne, death and burial. Getting an idea of Maya society from such information is a little like trying to reconstruct the society of 19th-century England by analyzing the gravestones in Westminster Abbey.

Other inscriptions are a bit more bountiful, describing something of the ritual life of the elite, including the ball game played by all Mesoamerican peoples, which is still poorly understood by scholars. A number of fascinating texts accompany stone "snapshots" of richly clad ballplayers in action. Usually two men are shown in competition bouncing a large rubber ball back and forth. It is a pity that the accompanying texts say nothing about the rules or scores of the games. Yet they do inform us that the rulers themselves sometimes played the game; often the ruler is called *ah pits*, or "ballplayer."

Of the special rituals described in the texts, the most common were personal rites of bloodletting and incense burning in which members of the nobility offered their most precious commodity—their blood—to the gods in exchange for divine favor and sustenance. These rituals were frequently carried out at such critical times as an enthronement, the designation of an heir or the celebration of a particular calendrical cycle. Although the rituals were significant—and were complemented by others in which the blood of captives was shed—it would be wrong to put too much emphasis on the bloodletting rites or to form the impression that Maya society was dominated by the shedding of blood.

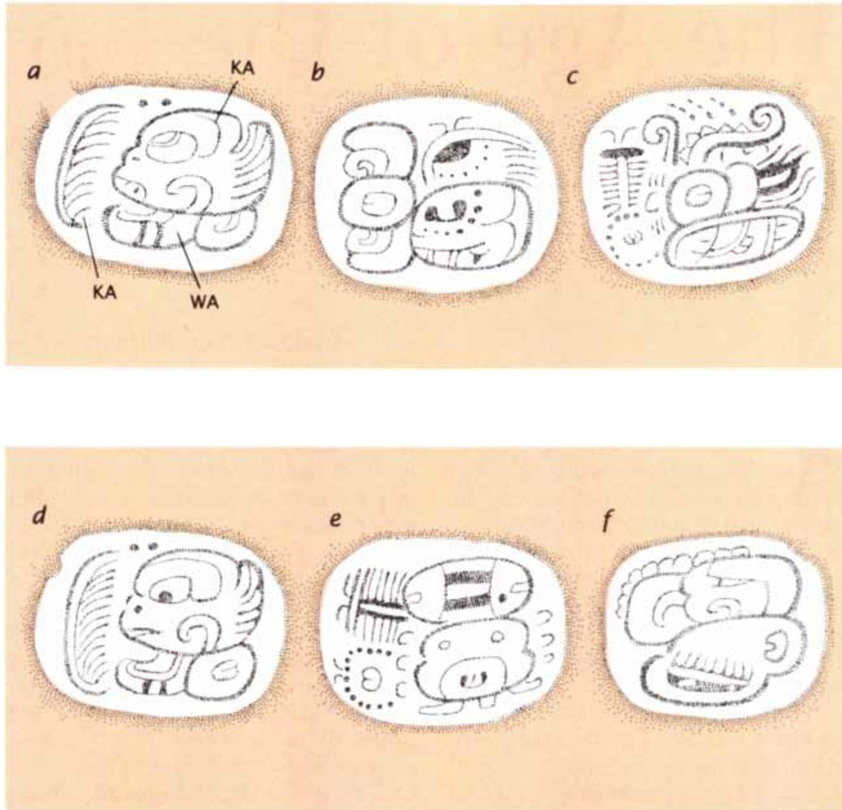


CACAO POT was discovered in a rich royal burial at Río Azul in Guatemala in 1984. The arching handle has a lock-top feature: it must be rotated slightly to remove the lid. The vessel is painted with coats of stucco that give the appearance of a jaguar pelt. Each of the 15 oval medallions originally was painted with a hieroglyph. In the pot were crusted remains of a liquid that proved on chemical analysis to be cacao.

Nevertheless, warfare between city-states did occupy a position of some importance. The taking of prisoners seems to have been a royal duty and a source of pride: one king from Yaxchilan is almost always referred to in the inscriptions as "he of 20 prisoners." In some instances, warfare seems to have led to great volatility in the control of territory. One area of the Maya lowlands where there are extensive records of warfare is the Lake Petexbatun region of Guatemala. Inscriptions there indicate that within a span of 40 years relations between certain city-states shifted from friendly to belligerent and back again. During the same period the polity centered at Dos Pilas in the Petexbatun region waged several raids, expanded rapidly and just as quickly lost its grip on the conquered land and shrank back to its former size.

The records from Dos Pilas suggest that Maya warfare was serious business. Inscriptions from other areas, however, imply that outright conquest was not always the aim of warfare and that ritual may have played a large part in these endeavors. Just how serious Maya warfare was in terms of territorial conquest may not be understood for some time. One recent discovery is, however, helping to clarify the understanding of Maya geopolitics: the discovery of precise place-names.

Place-name glyphs are well known from other Mesoamerican writing systems, but until now knowledge of places in the Maya area has been rather vague. The notion of Emblem Glyphs put forward by Berlin was a great advance, but it was not conclusive. Although Berlin believed that the Emblem Glyphs were of geographic significance, he could not be sure whether they referred to dynasties, to individual sites or to polities more inclusive than a single site. Following Peter Mathews of the University of Calgary, the majority of scholars now think Emblem Glyphs refer to polities



INSCRIPTION FROM CACAO POT has been partially deciphered by one of the authors (Stuart). The six glyphs shown here encircle the lid of the pot. Glyphs *a* and *d* spell the Maya word *ka-ka-w(a),* from which the English "cacao" is derived. Glyph *b* probably means "his vase." The paired references to cacao may indicate that two different types of drink were served in the vessel. The text continues on the body of the pot, giving the owner's name and perhaps his connection with a person of high rank.

or city-states larger than one specific site. Some of the most interesting current work on the Maya is aimed at understanding the exact role of these inclusive units in the framework of Maya geopolitics.

Within these polities, specific sites also had proper names that were expressed by glyphs in the inscriptions, as the authors of this article showed in 1986. Some of the specific names refer to natural features associated with a particular site. Aguateca, also in the Petexbatun region, was called by the ancient Maya *K'inich Wits,* or "Sun-Faced Hill." The name includes a "hill" sign split at the top, which is appropriate for a city atop a hill bifurcated by a chasm up to 50 meters deep.

Perhaps the largest number of specific names is found at the famous site of Tikal in Guatemala. Many places at Tikal are named in the inscriptions there, and although not all of the names have been translated, most of them probably refer to individual buildings or to building complexes. The names given to such structures by the Maya provide some insight into how they were conceived of by their builders. Some funerary pyramids, for example, have names that include the word *wits* ("hill"), which suggests that the Maya thought of them as artificial mountains. The inscribed upright stone stelae that dotted the Maya cityscape are frequently referred to as "plant stones" or "tree stones." In a literal sense, then, the Maya conceived of their cities as being composed of mountains and trees.

For the Maya the naming of things did not stop with buildings or monuments. Jade ornaments, stone vessels, ceramics, shell trumpets and other items were labeled with their names and those of their owners, as has recently become clear. Of these inscriptions, the largest group is found on ceramic vessels. In 1971 Michael Coe of Yale University announced that most of these texts on pottery included a string of glyphs in common, which he called the Primary Standard Sequence. Based on his examination of the accompanying painted scenes (which were largely mythological), Coe proposed that the Primary Standard Sequence was related to the *Popol Vuh,* the Maya creation epic that survived the Spanish conquest.

Quite recently, however, many of the signs in the texts on the ceramic vessels have been deciphered phonetically and semantically, and a very different interpretation has emerged. One common glyph has been translated as "painted," another as "his dish" (or "pot") and yet another as "cacao." The name of the vessel's owner customarily follows. This partial decipherment enables us to suggest that the inscriptions on the pots actually describe their ownership, function and contents. On a small scale this new interpretation reflects the larger movement away from the mythical and impersonal and toward the historical and personal that has taken place as the decipherment of Maya writing has proceeded.

During the past decade progress in decipherment has been impressive. As a result much new information has been gained about Maya history, language, social and political organization, and ritual life. Yet many inscriptions remain undeciphered, and nearly all texts contain hieroglyphs that cannot be interpreted completely. In some frustrating cases, signs can be deciphered phonetically, but their meanings are probably lost forever because the words do not exist in current Maya speech or in any extant dictionary. Other glyphs, however, will no doubt yield to future efforts at decipherment. Further surprises surely lie waiting in the unknown part of the most profuse and elaborate of pre-Columbian scripts.

FURTHER READING

MAYA HIEROGLYPHIC WRITING: AN INTRODUCTION. J. Eric S. Thompson. University of Oklahoma Press, 1971.

DECIPHERING THE MAYA SCRIPT. David Humiston Kelley. University of Texas Press, 1976.

EMBLEM AND STATE IN THE CLASSIC MAYA LOWLANDS: AN EPIGRAPHIC APPROACH TO TERRITORIAL ORGANIZATION. Joyce Marcus. Dumbarton Oaks, 1976.

THE MAYA, fourth edition. Michael D. Coe. Thames and Hudson, 1987.

MAYA ICONOGRAPHY. Edited by Elizabeth P. Benson and Gillett G. Griffin. Princeton University Press, 1988.

MAYA GLYPHS. S. D. Houston. University of California Press/British Museum, 1989.

The Age-of-the-Earth Debate

The controversy, which embroiled Archbishop Ussher, James Hutton, Lord Kelvin, Ernest Rutherford, Bertram Boltwood and Arthur Holmes, has "aged" the earth 4.5 billion years during the past three centuries

by Lawrence Badash

As the sun's first rays of thermonuclear light blazed across the galaxy 4.5 billion years ago, the primal earth emerged from a spinning, turbulent cloud of gas, dust and planetoids that surrounded the new star. During the next 700 million years the cloud settled into a more tranquil solar system, and the sun's third planet began to solidify.

On these figures for the age of the earth rest all of geology and evolution. Indeed, they seem to be part of humankind's permanent store of facts. Yet this chronological structure is quite new. In fact, two earlier estimates have toppled during the past 150 years as the descriptive sciences of biology and geology deferred to the more exact science of physics.

The first estimate fell during the 19th century. To the great displeasure of Charles Darwin and the geologists of the period, the physicist William Thomson (later Lord Kelvin) performed a seemingly flawless calculation to show that the earth had not existed throughout eternity, as many thought then, but had formed 100 million years ago. That chronology collapsed at the turn of the century, when the advent of radioactive dating techniques showed the earth's age to be a few billion years. After a fierce debate between geologists and physicists, radioactive dating prevailed. Above all, the age-of-the-earth controversy illustrates that emotion, intuition and vested interests can direct the course of science almost as much as logic and experimentation.

LAWRENCE BADASH is professor of history of science at the University of California, Santa Barbara. He was educated at Rensselaer Polytechnic Institute and Yale University, where he received his doctorate in 1964. His primary research interests are the role of scientists in the nuclear-arms race and the development of radioactivity and nuclear physics. In his spare time Badash has backpacked through the High Sierra in California, the Himalayas in Nepal and the Alps of Europe; he is a member of a mountain search-and-rescue team.

Intuitively, one might think questions about the earth's age were as timeless as speculation about the structure of the universe and our role in it. Actually, many of the earliest civilizations treated the earth's creation as part of the question of the origin of the universe. The resulting cosmologies tended to be cyclical. The Greeks, for example, believed natural history repeated itself perpetually. The Maya recorded 3114 before the common era (B.C.E.) as the year during which the universe had been most recently re-created. In the first century of the common era, many Han Chinese held a similar view. They believed the universe was destroyed and re-created every 23,639,040 years.

The Judeo-Christian tradition also combines the earth's and the universe's birth in a single event. The story of Genesis led scholars to calculate the number of human generations since Adam and Eve. In 1654 John Lightfoot refined Archbishop Ussher's famous calculation of the moment of creation to an ultimate degree of precision: October 26, 4004 B.C.E., at nine o'clock in the morning in Mesopotamia, according to the Julian calendar.

Mikhail V. Lomonosov was one of the first scientists to suggest (in the mid-18th century) that the earth formed independently of the rest of the universe; he set the interval at hundreds of thousands of years. In 1779 the Comte de Buffon tried to determine the age of the earth experimentally. He believed the earth was slowly cooling from an initial hot state, and he estimated that the earth was 75,000 years old by creating a small globe that resembled the earth's composition and then measuring the rate at which it cooled.

Lomonosov and Buffon were virtually alone in their rigorous pursuit of the absolute age of the earth. When other 18th-century naturalists pondered the question at all, they either placed everything in the hands of the Creator or else supposed that the earth and its living things had simply taken a long time to reach their present condition. James Hutton characterized the long view in his classic *Theory of the Earth* in 1795: "We find no vestige of a beginning, no prospect of an end."

The chronology of geological periods did, however, intrigue Hutton's contemporaries. They inferred that the successive strata of rock and soil at a particular site represent the order in which the layers formed. In the 1790's William Smith built on this perception: two layers from different sites could be regarded as equivalent in age if they contained the same fossils. Extrapolating from these ideas, the naturalists began to chronicle the strata and to estimate the duration of geological periods. Their estimates varied widely, since they could only make crude guesses about the time required to build up the layers.

In 1830 Charles Lyell gave such work a theoretical boost. Lyell insisted that rock formations and other geological features took shape, eroded and re-formed at a constant rate throughout time. Virtually none of the naturalists applied Lyell's notion to calculate the age of the earth's features; the data on geological processes were just too meager. Lyell did, however, persuade many naturalists to become uniformitarians—that is, they rejected the idea that there had been catastrophic geological change or a rapidly forming, young earth. After all, evidence from stones and bones suggested that each geological period lasted for many years, perhaps even hundreds of millions of years, and the age of the earth had to be several times that.

Therefore, the naturalists were startled when Lord Kelvin (then the physicist William Thomson from Glasgow) determined in 1862 that the earth had formed somewhere between 20 and 400 million years ago. Thomson rejected uniformitarianism as unprovable. He and many other physicists of the day believed the earth was originally molten; its surface had cooled and solidified, but the core remained hot. The deeper one descends into the earth, they noted, the higher the temperature.

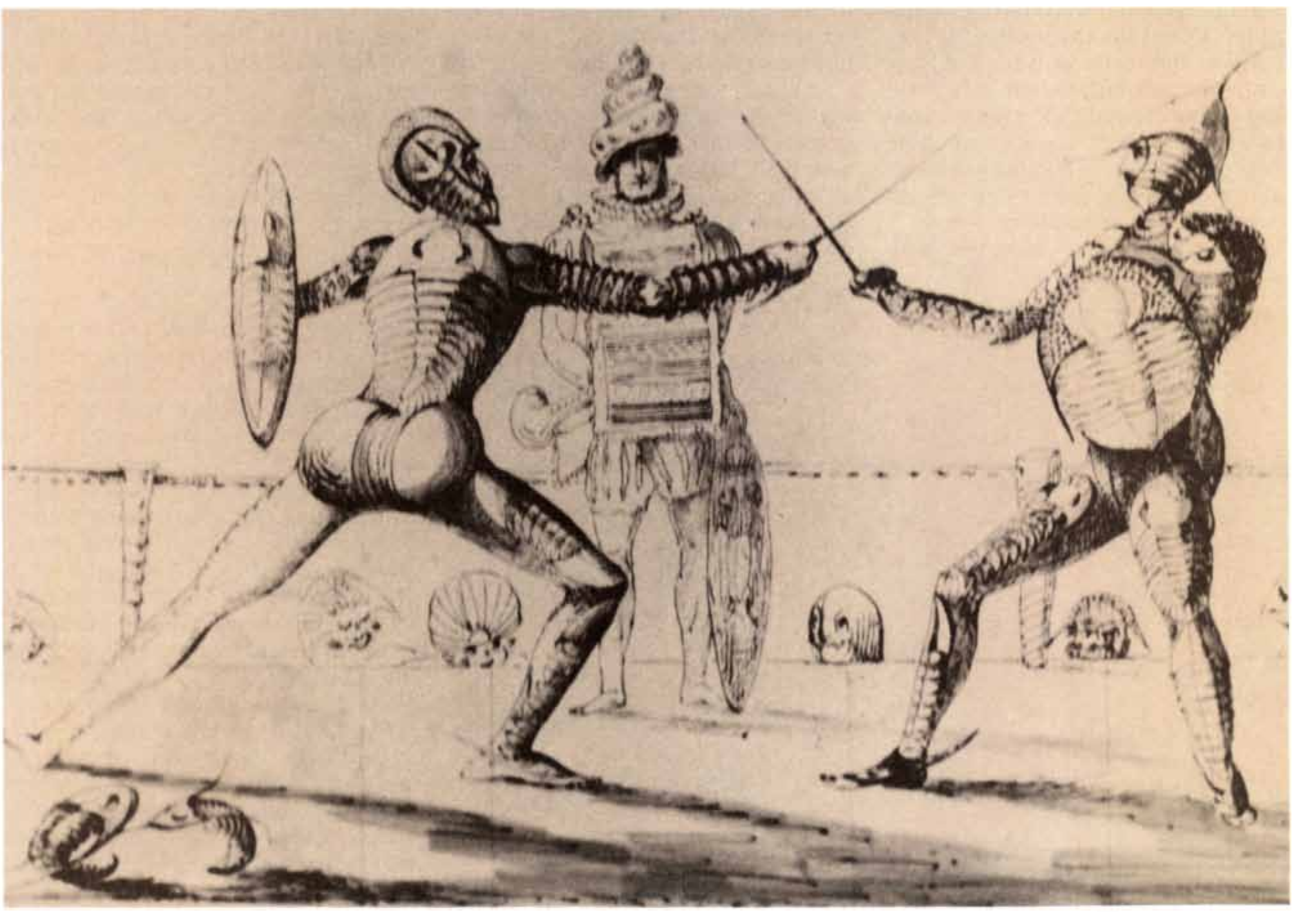
To derive the earth's age, Thomson calculated how long the earth required to cool from its primordial to its present state. He conjectured that the gravitational contraction that formed the earth had generated all of the earth's heat (except for a small contribution from the sun). Then he investigated how well the earth conducts heat and how much heat is necessary for it to melt or to raise its temperature by a certain amount. He knew that the earth had cooled steadily as energy radiated into the cold vacuum of space, according to the second law of thermodynamics. Using Jean-Baptiste-Joseph Fourier's theory of heat conduction, he predicted how the earth's temperature distribution might have evolved [see "The Fourier Transform," by Ronald N. Bracewell; SCIENTIFIC AMERICAN, June]. He corroborated his calculations by accounting for heat from the sun and the effects of tidal friction. In time he refined his estimate of geological history to from 20 to 40 million years.

Thomson's work distressed geologists, who were comfortable with the idea of unlimited time. They resented this audacious physicist who meddled in their field, but they could not fashion a counterargument, and they produced few papers on geochronology.

Thomson's calculation seemed unassailable on the grounds of logic and physics. His conclusion eventually proved to be inaccurate by a wide margin. Still, Thomson had instigated a conceptual coup d'état: qualitative geochronology was overthrown in favor of quantitative methods. Until the end of the century, Thomson's estimates were the standards against which all others were compared.

Thomson's result shocked biologists just as much as it surprised geologists. Darwin regarded Thomson as an "odious spectre" whose chronology was one of the shy naturalist's "sorest troubles." Darwin and other biologists had postulated that complex organisms would require much more than 40 million years to evolve. But neither living nor fossilized organisms offered a basis for an independent calculation of evolutionary time. The biological calendar ultimately relied on geology.

Thomas H. Huxley, a strong supporter of Darwin, attacked Thomson's most vulnerable position. Huxley's view epitomized the disdain that geologists of the late Victorian period felt for the physical sciences and the reluctant respect the workers held for quantifiable data. In his presidential address to the Geological Society of London in 1869, Huxley argued that



ARMED with Paleozoic fossils, stratigrapher Richard Griffith parries the attack by paleontologist John Phillips. The cartoon, drawn in 1843, satirizes a debate over the relative age of rock layers that contained these fossils. Such chronologies helped to establish the earth's age. Geologist Roderick Murchison referees, wearing a tabard that shows strata from the Paleozoic era.

no modern geologist would insist on absolute uniformitarianism but that its principles could be applied. Then Huxley directed his rhetoric at Thomson. The admitted "accuracy of mathematical processes [must not be permitted to] throw a wholly inadmissible appearance of authority over the results, [for] pages of formulae will not get a definite result out of loose data." Perhaps, Huxley suggested, heat radiated from the earth more slowly than Thomson supposed. Thomson thought he had estimated conservatively, but he could not be certain of his values.

Thomson no longer battled alone, however. Both the American astronomer Simon Newcomb and the German physicist Hermann von Helmholtz calculated the time needed for a nebular cloud to condense gravitationally to the present size of the sun. Their independent results of 100 million years established an upper limit for the age of the earth (presuming that the earth did not exist before the sun). George H. Darwin, son of the famous Charles and professor of astronomy at the University of Cambridge, joined the discussion. He posited that the moon broke loose from a rapidly rotating molten earth and found that Thomson's original estimates corresponded well to the time that terrestrial tidal friction would require to brake the earth to its present 24-hour period of rotation.

A few geologists concurred with Thomson's estimate of the earth's lifetime. Even before Thomson, John Phillips, Smith's nephew and pupil, had claimed that the earth must have endured for 96 million years—a result calculated from the admittedly imprecise rate of strata formation from river-derived sediment. In 1868 Archibald Geikie, director of the Geological Survey of Scotland, looked at evidence of erosion and concluded that the earth was no older than 100 million years.

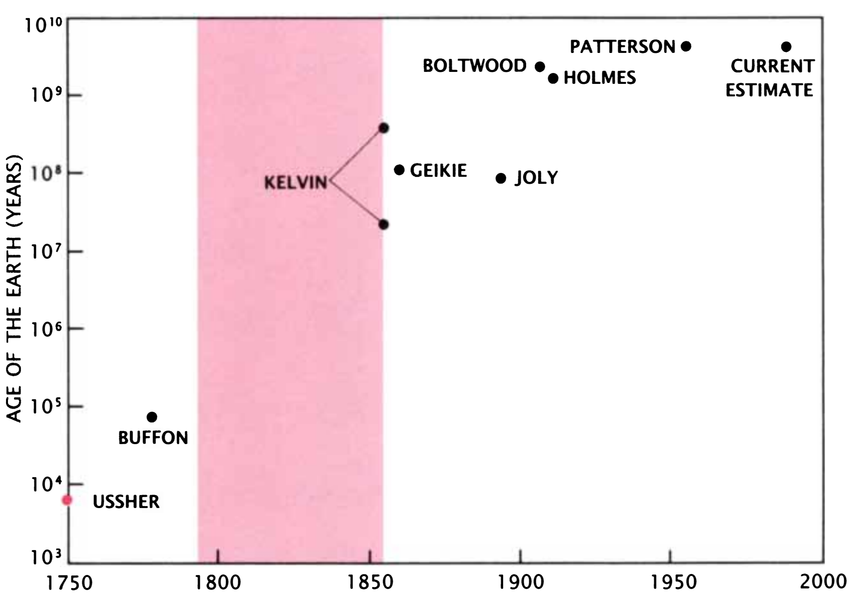
In 1899 John Joly of the University of Dublin devised the only truly new geological technique for measuring the earth's age. He maintained that all the salt in the oceans came from mineral deposits that had eroded and dissolved. He also proposed that the salt concentration in the oceans could not decline. Joly could therefore relate salinity to age. He obtained the best available figures for the quantity of water that flowed into the oceans each year and the amount of salt per volume of runoff. He then calculated the annual increment of salt. He multiplied the salinity of the ocean by its total volume and divided the product by the annual increase. Joly thus determined that the brackish sea developed over 80 to 90 million years.

At about the same time an increasing number of geologists swelled the consensus that the earth had formed less than 100 million years ago. Yet all attempts to measure the age of the earth rested on an assumption, an analogy or a best guess about the rate of change of geological processes.

Such assumptions created room for doubt. Some critics protested against the premise that only gravitational contraction explained the earth's or the sun's heat; another energy source might be possible. Some maintained that the earth had never been molten, whereas others suggested that its interior was still molten. (A liquid interior would conduct heat by convection—something that Thomson had not taken into account.) Still others questioned the data on erosion, sedimentation and salinity.

As the century drew to a close, geologists generally agreed that nearly 100 million years had passed since the earth was born. They did not, however, reconcile their differences with Thomson, who had recently been elevated to the peerage as Lord Kelvin for his scientific accomplishments. Employing his heat calculations, Kelvin was urging ever-shorter geological time scales, all the while high-handedly dismissing geological evidence.

By this time, however, geologists were wary of Kelvin's physical techniques. They had greater confidence (perhaps unwarranted) in their own methods than in the eminent physicist's collection of assumptions. After all, they had discovered several approaches to the chronology that gave concordant results. Geologists felt they had grandly completed their apprenticeship in the quantitative sciences after several decades of vigorous exploring, mapping, measuring and classifying.



GEOLOGISTS AND PHYSICISTS have advanced the earth's age from hundreds of human generations to billions of terrestrial revolutions. The red point marks the biblical estimates for the earth's age. Between 1795 and 1862 most geologists believed the earth had existed for eternity or at least a period beyond measurement.

Yet it was not long before physical scientists were once again treading on geologists' turf and calculating its age. This time the study of radioactivity gave momentum to the attack. In 1896 A. Henri Becquerel discovered the phenomenon; in 1898 Marie S. and Pierre Curie first detected the radioactive elements polonium and radium. Then in 1902 and 1903 Ernest Rutherford and Frederick Soddy explained the process of radioactivity in several papers. Radioactivity, they correctly stated, was the spontaneous transmutation of atoms of one chemical element into another.

At first, the radiation of alpha, beta and gamma rays was more important to geochronology than were the radioelements themselves. (It was later discovered that alpha particles are composed of two protons and two neutrons, just like the nucleus of a helium atom; beta particles are emitted electrons, and gamma rays are photons of electromagnetic radiation.) Earlier, in

1900, Rutherford and R. K. McClung of McGill University in Montreal showed that the various rays carry enormous amounts of energy. Their paper made little impression beyond the small community of physicists and chemists working on radioactivity.

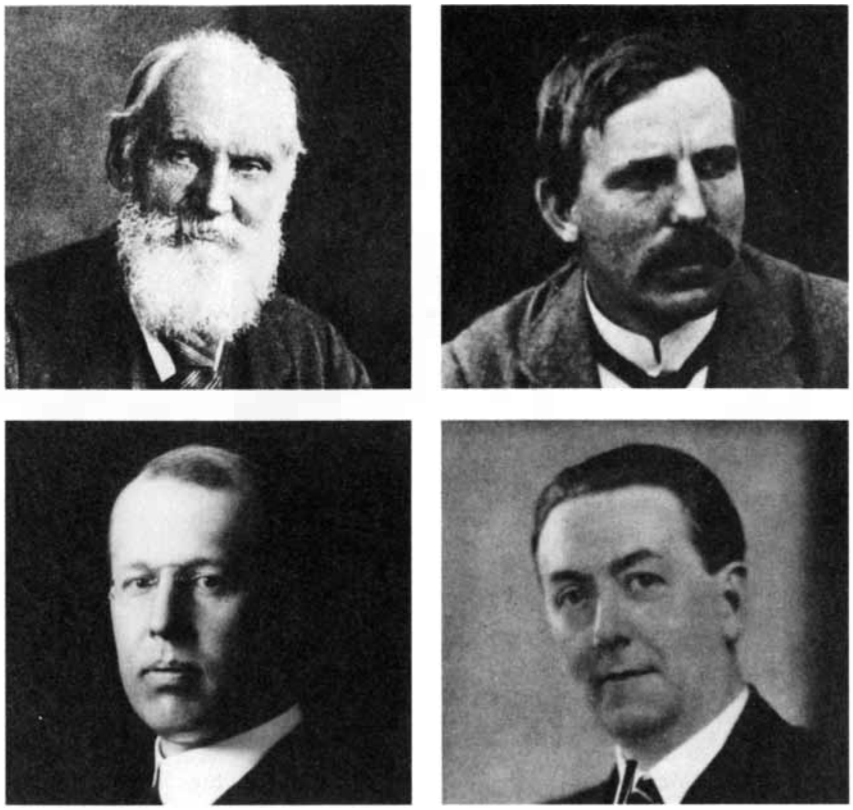
The reception was entirely different in 1903 when Pierre Curie and Albert Laborde announced that radium generates enough heat to melt more than its own weight in ice in an hour. Public interest was aroused over this apparently inexhaustible cornucopia of energy. Where did the energy come from? Rutherford and Howard T. Barnes discovered the source. They showed that the heat was proportional to the number of alpha particles radiated. These relatively massive particles were emitted at great velocity. If the particles collided with neighboring atoms, the particle's kinetic energy was transformed into heat.

Geologists immediately recognized that the relation between heat and radioactivity could significantly influence determinations of the age of the earth. Kelvin had assumed that the earth's heat came from either the sun or the original molten state of the earth. In both cases, gravitational contraction was the only source of energy. If the earth and sun contained quantities of radioactive materials sufficient to provide large amounts of heat, however, then this discovery could invalidate all chronologies that Kelvin had based on the earth's cooling.

In 1903 George Darwin and Joly were the first to make this very claim: radioactivity was at least partially responsible for the earth's and the sun's heat. But was there enough radioactive matter within the earth to make a measurable difference?

Part of the answer was at hand. Julius Elster and Hans F. Geitel, two schoolteachers in Wolfenbüttel, Germany, detected radioactivity in the air in 1901 and, soon after that, in the soil. Before long, many enthusiastic amateurs as well as professional scientists were finding radioactive rain, snow and groundwater—even radioactive mist at the base of Niagara Falls. Soon geologists had no doubt that radioactivity was widely distributed.

As for its concentration, Robert J. Strutt of the Imperial College in London found traces of radium in many rocks. Indeed, Strutt found too much radium for it to be distributed uniformly throughout the globe (without even considering the contribution from all the other radioelements). Its radioactivity alone could account for the earth's internal heat. His work suggested that geochronology could be extended by an indefinite time. He found no vestige of a cooling, no prospect of an age.



LORD KELVIN (*top left*) estimated that the earth could be a mere 20 million years old. Ernest Rutherford (*top right*) and Bertram B. Boltwood (*bottom left*) devised radioactive dating methods to determine that the earth formed more than a billion years ago. Arthur Holmes (*bottom right*) campaigned for the acceptance of these methods.

The scientific community responded with ambivalence. Joly and William J. Sollas of the University of Oxford worried that Strutt's work might overturn their own calculations demonstrating an age of about 100 million years. Kelvin's own feelings were divided: he privately acknowledged that his estimates had been superseded, but in public he remained contentious. Others were delighted to be liberated from Kelvin's earth age. It turned out that radioactivity not only loosened Kelvin's theoretical shackles but also held the key to determining the age of the earth.

During the first years of the century, scientific enthusiasm for radioelements replaced enthusiasm for radiation when Rutherford and Soddy proposed that radioactivity was actually spontaneous alchemy. A sample of a radioelement, they said, decays at a regular rate into a different chemical element. The rate of decay is expressed as a half-life: the time needed for half of the atoms of a given radioelement in a sample to change into a decay product.

Half-lives range from billions of years to millionths of a second. Uranium, thorium and radium have long half-lives and therefore exist in tangible quantities on earth, whereas those elements that have short half-lives have a transitory existence. Hence, the presence or absence of particular radioelements in rocks can imply an age; analysis of the quantities of the radioelements can reveal an absolute age.

The radioelements form distinct decay series: one radioelement decays into the next element in the series until a stable element is produced. The uranium-radium, uranium-actinium and thorium series were known or suspected in the early years of this century. The technique of radioactive dating of rocks developed from the study of radioelements and their decay series. Rutherford and the radiochemist Bertram B. Boltwood pioneered the work.

As a consulting chemist after his graduation from Yale University, Boltwood examined numerous ore samples, among them monazite, a miner-

CARTOGRAPHERS, who constructed the Hunt-Lenox Globe (*left*) in 1500, could only refer to the Bible to estimate the age of the earth. They would have miscalculated by a factor of more than 800,000. Astronauts, who photographed the earth from space (*right*), could rely on measurements of radioactive elements to date the earth's creation back 4.5 billion years.

al containing uranium and thorium. When the charismatic Rutherford lectured at Yale in 1904, Boltwood's curiosity about radioelements became a passion, and he began to document the relations among the elements in the decay series.

Later that year, Rutherford suggested a way to determine the age of the earth from measurements of helium in minerals. Rutherford then believed (and in 1908 proved) helium is not a product of any particular decay series but is formed in all the series when two electrons bond to an alpha particle. Sir William Ramsay and Soddy at University College in London had just discovered the rate at which radium produced helium. If the Ramsay-Soddy rate was accurate and no helium escaped from the mineral from the time of its formation—both great leaps of faith—the amount of helium would determine the age of the sample. Rutherford could boast an age of 40 million years for a fergusonite rock he owned.

Boltwood, on the other hand, thought to look for the end products of the decay series. The amount of an end product would increase over the years as the radioelements decayed. It was already known that radium was a product of the uranium series; in 1905 Boltwood pointed to lead as the final product. The uranium-to-lead hypothesis received additional support from Rutherford. He argued that if uranium decayed to radium and if radium (then thought to have an atomic weight of 225) and its daughter products then emitted five alpha particles (which each have an atomic weight of four), the decay would yield an element of an atomic weight of 205—not far from lead's accepted value of 206.9.

Boltwood credited Rutherford for suggesting the lead method of dating ancient rocks, but it was the chemist who demonstrated its feasibility. By the end of 1905 he had calculated ages ranging from 92 to 570 million years for 26 different mineral samples. Fortunately for the reputation of the new technique, these results remained unpublished. Boltwood's radium-to-uranium ratio was inaccurate both because Rutherford's scale for measuring quantities of radium was badly calibrated and because the half-life of radium was revised several times during 1905 and 1906. A rock's age rested critically on both these values.

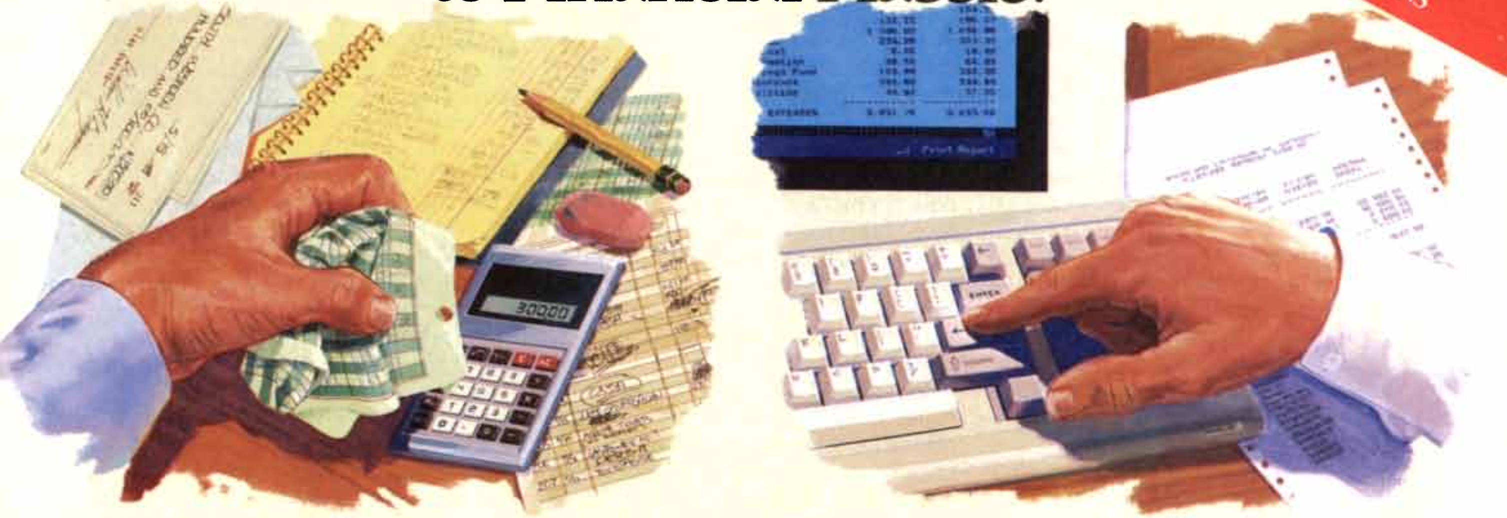
When Boltwood published his work in 1907, he reported a striking constancy in the lead-to-uranium ratios for minerals from the same rock layer, which were presumably of the same geological age. He also observed that the amount of lead in a mineral increased as the relative age of the mineral increased. Minerals from which lead had apparently been leached gave lower ratios than did other minerals from the same layer. All this fit together well. Boltwood could find, however, no constancy in lead-to-thorium ratios from several minerals; the end product of thorium remained a mystery. He was inclined, therefore, to ignore lead-to-thorium ratios: an error that affected his measurements of minerals that contained both uranium and thorium.

To determine the absolute age of minerals, Boltwood examined the uranium-radium decay series. The latest value for the half-life of radium was 2,600 years, which Rutherford had deduced from the number of alpha particles emitted from radium each second. (The figure accepted today is 1,620 years.) Given that the decay of radioactive materials is exponential, the fraction of radium decaying in one year would be 270,000 parts per billion, based on Rutherford's half-life. Rutherford and Boltwood found that almost all rocks contained 380 parts of radium per billion parts of uranium. Thus, the fraction of radium decaying each year multiplied by the fraction of radium in uranium yields one part of radium decaying each year for every 10 billion parts of uranium.

Boltwood correctly assumed that the decay series of the rocks he collected were in an equilibrium state. The uranium-to-lead series, for instance, is in equilibrium when the number of uranium atoms decaying per unit of time is equal to the number of radium atoms decaying, or lead atoms forming, in that time. To maintain this equilibrium, radioelements that have long half-lives must exist in greater quantities than those that

Say Goodbye to Financial Hassle!

NEW VERSIONS



So Long, Tedium   Hello, Quicken

Why waste time doing your finances? Quicken slashes time and effort *every* month…perfectly organizes your finances…and provides instant insight into your money matters.

Breeze Through Financial Chores

Quicken takes the drudgery out of routine finances. It automates everything from bill paying to bookkeeping…reconciliation to reporting. Quicken does in minutes what used to take hours.

Quicken actually learns your regular bills and reminds you when to pay them. Just confirm payment amounts, and *in seconds,* Quicken performs 14 bookkeeping steps, categorizes expenses, and hands you printed checks addressed for mailing. Or, pay your bills using CheckFree,™ the nation's electronic payment system. And, Quicken keeps perfect records of all your handwritten checks.

A Bigger Picture with Smaller Effort

Quicken automatically organizes your finances. Imagine always knowing how much money you have…and exactly where it's going. At the press of a button, Quicken generates insightful reports such as Income and Expense, Cash Flow, Budget vs. Actual, Balance Sheet/Net Worth, Job Costing, and *more.*

Relax at Tax Time

Stop dreading tax time. Quicken keeps important tax data at your fingertips automatically. In *seconds,* you can print an itemized list supporting your deductions. Plus, Quicken links directly to top selling tax software like **Turbo Tax®** and **J.K. Lasser's Your Income Tax.®**

NEW IBM® Version
- New credit card, cash, asset, and liability accounts
- New reports show data from all accounts
- New Tax Summary and Net Worth/Balance Sheet reports
- Variable monthly budget reports
- Links with CheckFree,™ the nation's electronic payment system

NEW Macintosh® Version
- Far more flexible reports, with multi-column capability for trend tracking

Both New Versions
- New wallet-size checks
- New Canadian dates and tax categories

The World's #1 Best Seller

Quicken is the best selling business and personal financial software in the world. Bill Howard of **PC Magazine** (December 27, 1988) says, **"Quicken is about as good as software gets—powerful, affordable and easy to use."** Broadway playwright ("Mame") Robert E. Lee remarks, **"This astonishingly simple program…is truly true to its name. Indeed, it does quicken the time required for routine bookkeeping."**

INFO WORLD 1988 PRODUCT OF THE YEAR

PC MAGAZINE Best of 1988 AWARD

No Computerese or Accounting Mumbo Jumbo

Quicken is so easy, you're up and running in minutes. That's because the whole program looks and works just like the paper checkbook and register you've always used. Quicken doesn't change *how* you do your finances, only how *fast* you finish them.

Powerful Business Features

Over half of Quicken's customers use it for small business bookkeeping. Quicken produces financial statements such as Profit and Loss, Balance Sheets and other useful management reports. It forecasts your cash flow, and tracks A/R, A/P, payroll, jobs, clients, properties, and more.

SPECIFICATIONS

Feature sets vary according to machine type. Net Worth/Balance Sheet reports, available for the IBM, are not available for Mac and Apple versions.

IBM: All IBM® PC, XT,® AT,® PS/2,® and compatibles with 320K RAM and DOS 2.0 or higher. Exports in ASCII and Lotus® formats.

Mac: All with 512K RAM. Exports to SYLK, Text, and HyperCard™ formats.

Apple: IIGS, IIc, IIc Plus, and IIe with 128K RAM. Exports to AppleWorks® for budget, custom, P&L, and multi-column reports.

All versions: Works with all monitors and all printers. Unlimited number of bank accounts and transactions (subject to available disk space). Maximum amounts to $9,999,999.99.

Other: Unlimited FREE technical support. Not copy-protected. Checks are pre-approved by all financial institutions in the U.S. and Canada. Sample checks and order form are enclosed in Quicken package.

CheckFree Electronic Bill Payment: IBM Version only. Pays all bills. Works with all U.S. financial institutions. Modem and software required. A service of CheckFree Corporation.

stores.

Try Quicken FREE for 30 Days

Order from the coupon below, and if you're not completely satisfied for *any* reason, simply return it for a full, no-questions-asked refund. So, why wait? Say good riddance to financial hassle, and order Quicken *today.*

Send in the coupon, or call 1-800-624-8742 U.S. and Canada.

$59.95

Yes, I want to say goodbye to financial hassle and try Quicken. If I'm not completely satisfied, I may return it within 30 days for a full refund. Quicken, U.S. $59.95, includes free shipping by air. California residents, please add sales tax. For coupon orders, allow 10 days for delivery (longer to Canada).

Name ______

Address ______

City ______ State ______ Zip ______

Daytime Phone ( ) ______

☐ IBM and compatibles ☐ 5¼" disk ☐ 3½" disk
☐ Apple II
☐ Macintosh

Preferred payment method (check one)
☐ Check enclosed (no purchase orders or COD, please)
☐ Visa ☐ MasterCard ☐ American Express
Card No ______ Exp. Date ______

Signature ______

Mail to:
Intuit
540 University Ave.
Palo Alto, CA 94301

Or, call: 1-800-624-8742 in U.S. and Canada for charge card orders.

Current Quicken users:
To upgrade, call the 800 number above.

G-SA08

have short half-lives. (Although the supply of uranium will slowly decrease over time, Boltwood realized that the amount lost is negligible.)

Boltwood deduced that if one part of radium decays each year for every 10 billion parts of uranium, then one part of lead forms each year for every 10 billion parts of uranium. Boltwood expressed this relation in a formula: the age of the rock equals 10 billion years multiplied by the ratio of lead to uranium. He then calculated that a sample of uraninite, which had a ratio of .041, was 410 million years old and a sample of thorianite, which had a ratio of .22, had formed 2.2 billion years ago.

Actually, when the accurate value for the half-life of radium was applied, the age of Boltwood's samples was found to range from about 250 million to 1.3 billion years. Even with this correction, his thorianite measurement was invalid because the decay of thorium contributed some lead in addition to the lead that derived from uranium. Nonetheless, these results were spectacular: they demonstrated that the earth was about a billion years old.

Oddly, this enormous accomplishment was met with indifference. Although Boltwood's paper appeared in America's foremost geological journal, no one was inspired to duplicate or extend his work on the lead method. Nor did Boltwood's result sway geologists' opinion that the significance of radioactivity was overrated. They not only discounted the heating effect of radioactive decay on the earth but also "refined" their geological and physical data to show that Kelvin's range of time was correct after all!

Boltwood wrote no more on the lead dating method. He returned to the study of decay series and discovered ionium, the immediate parent of radium. Rutherford retained a light hold on the age-of-the-earth topic, publishing about one paper a decade—hardly the mark of a consuming interest. Meanwhile, Strutt refined the helium method until 1910, when he too departed for greener research pastures.

Strutt left a legacy, however. He had sparked an interest in geochronology in a young English geology student, Arthur Holmes, who kept the subject alive almost single-handedly. Indeed, Holmes ultimately forced geologists to accept radioactive dating in the course of his long career in industry and at the Universities of Durham and Edinburgh. Until 1930, however, Holmes and Joly were the only geologists who were skilled in the dating technique, and Joly, moreover, doubted its accuracy.

Holmes did not. He also considered the lead method to be more reliable than the helium technique. In 1911 he examined many rock samples and calculated that the most ancient was 1.6 billion years old. He maintained (with more faith than justification) that his samples had contained no lead when they were formed, that all the lead came from the decay of uranium and that external mechanisms had not removed or added any lead or uranium.

Two years later, however, his critics could crow in the light of two new advances. The first was the discovery of isotopes: atoms that have the same chemical properties but different atomic weights because the number of neutrons varies. Lead, for example, has a nucleus that contains 82 protons and can have an atomic weight from 195 to 214. The second advance was the discovery of the physical laws that specify the decay products of each radioelement. These laws indicated that the thorium series did after all end in a particular isotope of lead.

Although for many earth scientists these new discoveries made radioactive dating seem more difficult and unreliable, Holmes forged ahead, publishing in the years before and after World War I a steady stream of papers on geochronology. He incorporated information about isotopes into his work and sharpened his results. Although his success wore down overt resistance to radioactive dating, the method gained little support.

An exception was Joseph Barrell, a professor of geology at Yale, who in 1917 reinterpreted geological history to conform with the results of radioactive dating. Barrell emphasized that geological processes vary in intensity in a cyclical rather than a uniform fashion. Thus, current rates of geological change could not, as uniformitarians claim, be a guide to the past.

Finally, resistance began to falter. By 1921, at a meeting of the British Association for the Advancement of Science, the speakers, representing geologists, botanists, zoologists, mathematicians and physicists, seemed to agree that the earth was a few billion years old and that radioactive and geological dating techniques could be reconciled. But no plan was drafted for reconciliation. Not surprisingly, the old guard remained skeptical. Sollas would accept no age for the earth greater than 100 million years. "Geologists," he said, "are not greatly concerned over the period which physicists may concede to them; they do not much care whether it is long or—in moderation—short, but they do desire to make reasonably certain that it is one which they can safely trust before committing themselves to the reconstruction of their science, should that prove to be necessary."

The battle was won finally in 1926 when in the U.S. the National Research Council of the National Academy of Sciences appointed a committee to examine the status of the age-of-the-earth problem. Holmes, as one of the few experts on the subject, was a committee member and wrote almost 70 percent of the report. The committee agreed unanimously that radioactivity provided the only reliable geological time scale. The report presented an overwhelming amount of clear and detailed evidence. The constants of radioactivity were firmly established, lead isotopes were easily incorporated into the calculations, and the mineral samples were carefully chosen to ensure that decay products had not been lost over time. The radioactive dating methods pioneered by Rutherford and Boltwood and enhanced by Holmes had at last received the blessing of geologists. Not only had they found a vestige of a beginning, but they also had a prospect for dating all of geological history.

During the past six decades, application of the lead dating method has become more and more sophisticated, and current techniques reveal that the oldest rocks on the earth were formed as much as 3.8 billion years ago. This would date the minimum age of the earth's solid crust but not necessarily the period when the spiral cloud of gas and dust condensed to form the solar system. In 1955 Clair Patterson of the California Institute of Technology and his colleagues first determined the age of the solar system by dating meteorites. The most recent measurements of meteorites place the age of the primal earth at 4.5 billion years.

FURTHER READING

GENESIS AND GEOLOGY. Charles Coulston Gillispie. Harper, 1951.

RUTHERFORD, BOLTWOOD, AND THE AGE OF THE EARTH: THE ORIGIN OF RADIOACTIVE DATING TECHNIQUES. Lawrence Badash in *Proceedings of the American Philosophical Society*, Vol. 112, No. 3, pages 157-169; June, 1968.

LORD KELVIN AND THE AGE OF THE EARTH. Joe D. Burchfield. Science History Publications, 1975.

THE AMATEUR SCIENTIST

How to analyze the shock waves that sweep through expressway traffic

by Jearl Walker

In light traffic the driver on an expressway can maintain a fairly steady speed—even let a car equipped with cruise control do so all by itself. In moderate to heavy traffic, on the other hand, the car's speed often needs to be adjusted if efficient progress is to be made and collisions avoided. Usually the procedure is merely annoying; the adjustments do not require much concentration. Sometimes, though, the cars ahead of you slow down so suddenly that you need to stand on the brake pedal, quickening the pulse and bringing the task of driving sharply to the conscious level. What conditions produce such a dangerous situation?

The movement of cars along a freeway may seem to involve so many variables as to defy analysis, but since the 1950's several methods of examination have been introduced and gradually refined. The methods liken the flow of cars to the flow of a fluid, focusing on the average properties of the traffic and ignoring the detailed and random fluctuations.

One early study was published in 1955 by Michael J. Lighthill and Gerald B. Whitham of the University of Manchester. They proposed that small variations in the flow sweep through the traffic as "kinematic waves." If waves traveling at much different speeds happen to meet, a "shock wave" develops, and it is at the site of the shock wave that drivers must suddenly brake. Depending on circumstances, a shock wave can travel in the direction of traffic (downstream) or in the opposite direction (upstream), or it can even be stationary.

Two of the possibilities are shown in the illustration on the opposite page: two series of overhead views of cars moving along a single lane. The three columns on the left demonstrate a downstream shock wave, the three on the right an upstream shock wave. In each group the first column is an early view, and the next two columns are progressively later views; in each group speedy and widely spaced cars near the bottom catch up with a pack of slower-moving and more closely spaced cars near the top. The shock wave is at the car that has just been braked hard to avoid a collision with the trailing slow car in the pack. Notice that in the left-hand group the shock wave travels downstream, that is, each successive car goes a little farther before having to brake; in the right-hand group the shock wave travels upstream. After outlining the recipe developed by Lighthill and Whitham for studying these situations, I shall explain how you can put their analysis to the test with only a stopwatch, a video camera and a videotape player.

The analysis involves measurements that are tedious to make, but it leads to a simple and yet powerful technique for predicting how kinematic waves and shock waves develop and move. The measurements allow you to calculate the flow, concentration and mean speed of the traffic. First, two lines are mentally drawn across the lanes in which the traffic is moving in one direction. The lines are separated by some short distance *L*, perhaps 40 meters or so. For a certain time period *T*, say 60 seconds, you count the number of cars that move through *L*, and you also record the time each car takes to move through *L*. (That last measurement, of course, is impossible if you simply stand at the side of the road and watch the traffic, but the chore becomes feasible if you record the traffic on videotape, which can be reversed and rerun at will.)

The flow of traffic through *L* is defined as the number of cars you count divided by the time *T*. To calculate the concentration of cars, you add up their transit times through *L* and divide the sum by the product of *T* and *L*. To get the mean speed, you divide the flow by the concentration. (The equation can be rearranged into a form that will be useful later: the flow is equal to the product of the concentration and the mean speed.) For example, suppose that 20 cars pass through an *L* of 40 meters in 60 seconds and the sum of their transit times is 36 seconds. The flow through *L* is then 1/3 car per second, the concentration .015 car per meter and the mean speed 22.2 meters per second.

These results are, of course, only samples, because from moment to moment the number of cars and their transit times through *L* vary. If you repeat the measurements, the samples reveal more definitively how much traffic travels through *L*. I should particularly emphasize that the computed speed is only a mean. If the traffic is especially light or heavy, most cars travel at the mean speed, but for other concentrations none of them may travel at that speed. Some of the cars may even go considerably faster, particularly on a multiple-lane expressway, where an aggressive driver can (at some risk of collision) switch lanes when approaching a slower driver.

Imagine that you review a videotape of expressway traffic as it gradually builds from a low concentration to the greatest possible concentration: a traffic jam in which all the cars stop. At first the rare car that passes through *L* travels at the speed limit. Once the concentration increases somewhat, the separation of the cars decreases to the point where the drivers begin to "interact," that is, they reduce their speed because of proximity. The reduction may result from a concern for safety, each driver slowing because he (or she) senses that the car in front might suddenly stop. The reduction may also arise because the increased concentration lessens the possibility of a lane change, so that cars begin to pile up behind a slow driver. As the concentration continues to increase on the videotape, the mean speed continues to fall, for either or both reasons. At the limit, the concentration reaches its greatest value, and the cars are stationary, bumper-to-bumper.

What happens to the flow during the buildup in concentration? Recall that flow can be expressed as the product of the concentration and the mean speed and that the former increases at the expense of the latter. Initially the flow increases owing to the increase in concentration and in spite of the slowing of the cars. Later on the tendency is reversed: the flow decreases because of the slowing and in spite of the increase in concentration. The results can be displayed by a "flow graph" in which the flow is plotted against the concentration [*see upper illustration on next page*]. Notice that at some intermediate concentration the flow is maximum; such a state is rare for expressway traffic. Before rush hour the relevant part of the graph is on the left side. During rush hour the right side of the graph dominates. The transition between the two traffic states is usually rapid.

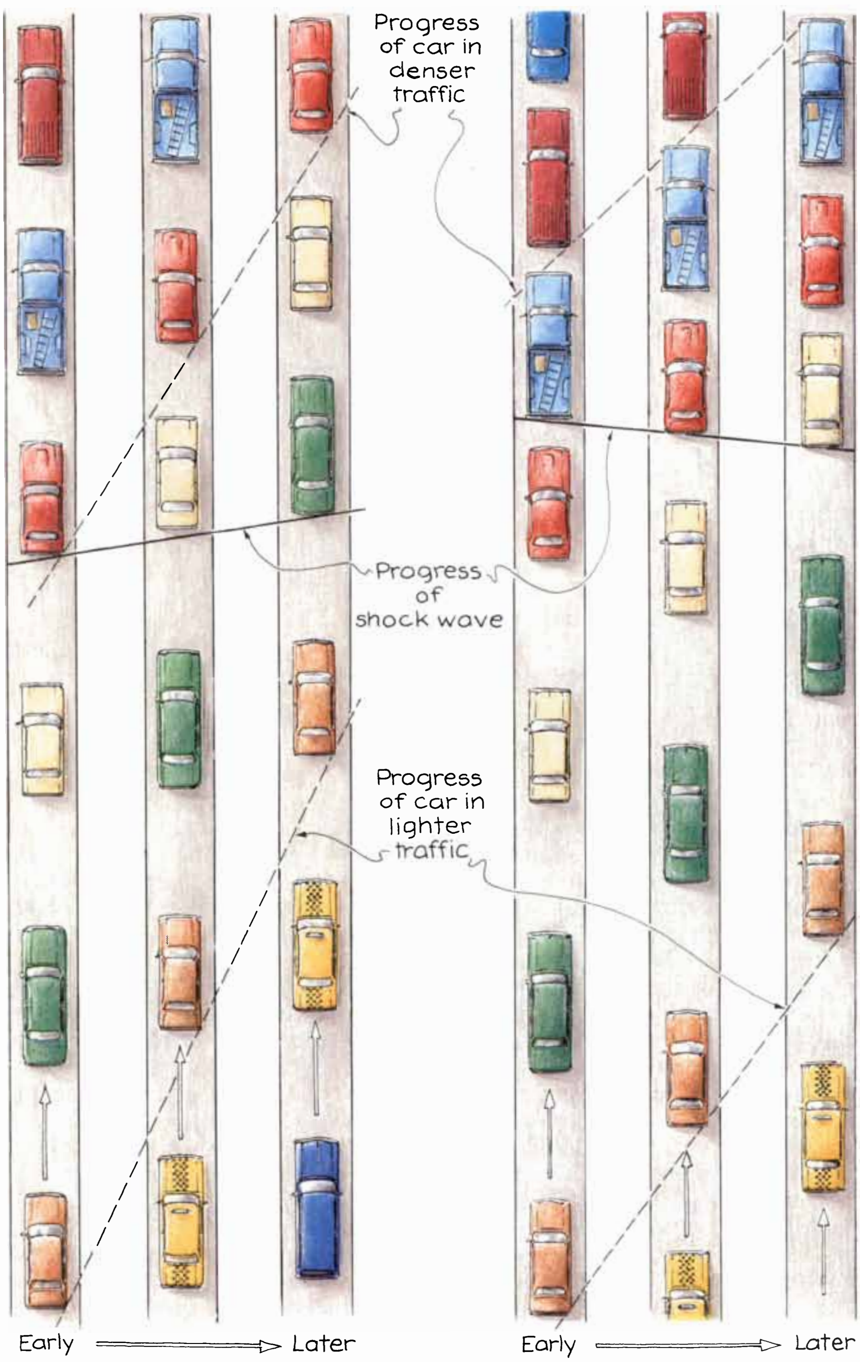
Once a flow graph has been constructed from observations and measurements, the mean speed of the cars for any given concentration can be derived. Draw a straight line from the origin of the graph to the point on the curve corresponding to the concentration in question. The slope of the line is the mean speed for that concentration. Note that if you first consider a point on the left side of the curve (such as *A*) and then consider a point farther to the right (such as *B*), the slope of the line decreases, as does the mean speed. If the point is at the extreme right of the curve, the line has zero slope, and the cars are frozen in place in a traffic jam.

So far the analysis provides little predictive capability, but Lighthill and Whitham noted a more powerful aspect of the graph: it shows how the kinematic waves travel through the traffic. The waves result from adjustments the drivers make in response to small variations in concentration. Suppose that you drive through moderate traffic. When your speed matches the speed of the car ahead of you and you do not want to pass, you tend to maintain your speed. If, however, the driver ahead of you slows down somewhat, the distance between your cars narrows and the concentration there increases, and so you too slow down. Your reaction is not instantaneous; it takes about a second. After another second, the driver behind you begins to slow down as well. Thereafter, the process of slowing travels back through the line of cars in your lane as a kinematic wave.

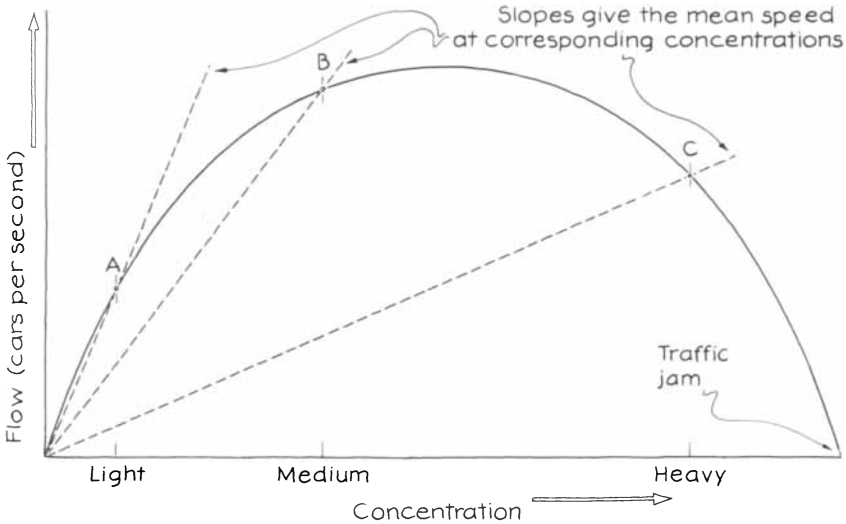
A similar wave is created if the driver in front of you speeds up a bit, increasing the space in front of you and decreasing the concentration. After a second or so, you too speed up. After another second, the driver behind you speeds up, and so on. In both examples the change in speed approximately offsets the change in concentration, and so the flow (which is the product of the two factors) remains constant. For this reason kinematic waves are said to be waves of constant flow: their existence keeps the flow constant in spite of inevitable small fluctuations in the speed of individual drivers.

How fast kinematic waves travel through the traffic depends on the overall, or average, concentration, and it can be determined from the flow graph [*see lower illustration on next page*]. Find the point on the curve corresponding to the concentration, and draw a tangent to the curve at that point. The velocity of the waves (with respect to the road or to an observer alongside the road) is equal to the slope of the tangent. If the point is on the left side of the graph, the slope is positive, which means that the waves travel downstream. If the point is on the right side, the slope is negative, and the waves travel upstream. When the point is at the top of the curve, the tangent has no slope, and the waves are stationary.

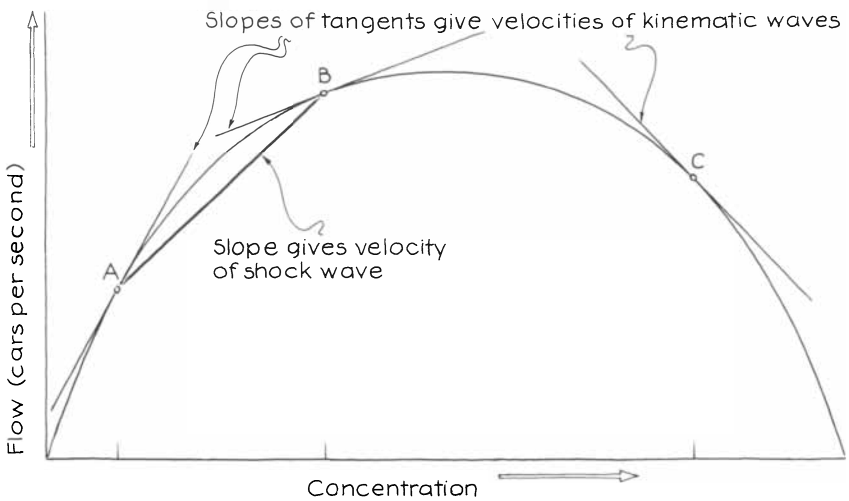
In all cases the waves travel more slowly than the mean speed of the



Cars in a single lane encounter downstream* (left) *and upstream* (right) *shock waves



A flow graph and the derivation of mean speeds



The velocities of waves

cars and so must travel back through the traffic. A wave is probably invisible to anyone watching the traffic from the roadside because the adjustments associated with it are so slight. It is invisible even when you drive through it; you perceive only the distance to the car in front of you and the adjustments you make in your own speed.

The information about the mean speed and the velocity of the kinematic waves becomes more useful when it is transferred from the flow graph to a "distance graph," in which distance along the expressway is plotted against time [*see illustration at left on page 100D*]. The lower part of the new graph pertains to the beginning of a stretch of highway, and the upper part pertains to a more distant segment. The left side reflects traffic conditions early in the period of observation, and the right side reflects conditions later on.

Suppose that the concentration and flow are the ones represented by *A* on the flow graph. Draw a straight line from the origin to *A*, and then draw on the distance graph a line that has the same slope. This second (*broken*) line represents the progress of a car that happens to move at the mean speed of the traffic. The car enters the picture at some early time and then moves steadily along the expressway as time passes. Additional lines can be added to represent other cars that enter the picture at other times and that move at the mean speed.

The information about kinematic waves can also be transferred between the graphs. Draw a tangent to *A* on the flow graph, and then, on the distance graph, draw a series of lines that have the same slope as the tangent. Their spacing is immaterial, but their slope reveals how the waves travel along the stretch of highway being studied. A wave that first appears at the near end of the highway gradually moves downstream (toward the top of the graph). You can do the same thing for the concentration associated with point *B* on the flow graph. In this case, the lines representing the mean speed and the kinematic waves all have less slope than in the preceding example.

If the concentration is dense, the relevant part of the flow graph might be point *C*. The mean speed is lower than it is in the previous examples, and the kinematic waves are now represented by lines that have a negative slope. The slope indicates that a wave begins far down the expressway and moves steadily toward the nearby segment.

The graphs help one to picture how a shock wave moves along an expressway. A wave might develop when something increases the concentration suddenly and thus requires the mean speed to decrease dramatically. For example, a shock wave can be created when a heavy truck pulls onto the expressway and travels slower than the prevailing mean speed. As the cars in the truck's lane approach the truck, they suddenly slow, and their concentration increases. The changes in speed and concentration travel back through the approaching traffic as a shock wave.

Suppose that before the truck appears the flow conditions are represented by *A* on the flow graph and that afterward they are represented by *B*. The slope of a line connecting *A* and *B* equals the speed of the shock wave. It is also about equal to the average of the slopes of the tangents at *A* and *B*. The line has a positive slope, which means the shock wave travels downstream. To follow its motion, transfer the line to a distance graph [*see illustration at right on page 100D*]. Complete the graph by adding lines to represent the kinematic waves and the travel of a car at mean speed for conditions associated with points *A* and *B* on the flow graph. In front of the shock wave (higher on the graph), the conditions are those associated with point *B*. Behind the shock wave (lower on the graph), the conditions are those associated with point *A*. Notice that the shock wave is the line along which the kinematic waves having one speed intersect with kinematic waves having a much different speed. In this example, the intersection produces a shock wave that moves downstream.

Suppose that you were to drive

SPECIAL ADVANCE ANNOUNCEMENT

Due to the expected high interest in this September 1989 single-topic issue and a limited print run, the publisher is offering a special opportunity to our readers to order extra or bulk copies in advance.

"Managing Planet Earth" will be a path-breaking issue that belongs on everyone's reading list and should be included in next year's curriculum. Share this vital information with your colleagues, students and anyone else for whom these pressing issues are a concern. Clip the Pass-Along Order Form below and bring it to their attention!

ORDER FORM

Yes, enter my order for ______copies of the September 1989 single-topic issue, "Managing Planet Earth", at $______ per copy. Total order $______.

*1 to 9 issues $3.95 per issue
*10 to 24 issues $3.75 per issue
*25 to 99 issues $3.50 per issue
*100 or more issues $3.20 per issue
Quantities limited.

PRICES INCLUDE SHIPPING (U.S. ONLY*)

❑ $________Payment enclosed
❑ Purchase order #________attached
❑ Charge my credit card (4 copy minimum):
❑ VISA ❑ Mastercard ❑ Access ❑ Eurocard
❑ Please Bill Me (4 copy minimum)

Card # ____________ Exp. Date ____________

Signature ____________

Name ____________ School ____________

Address ____________

City ____________ State ________ Zip ________

For more information, call Ms. Davis at 1-212-754-0539.
Expected shipment date Aug. 29, 1989.
Cut and mail this coupon to: 415 Madison Avenue, Dept. MPE, NY NY 10017
*Outside the U.S. add $1.00 per copy.

P9AA013

PASS-ALONG ORDER FORM

Yes, enter my order for ______copies of the September 1989 single-topic issue, "Managing Planet Earth", at $______ per copy. Total order $______.

*1 to 9 issues $3.95 per issue
*10 to 24 issues $3.75 per issue
*25 to 99 issues $3.50 per issue
*100 or more issues $3.20 per issue
Quantities limited.

PRICES INCLUDE SHIPPING (U.S. ONLY*)

❑ $________Payment enclosed
❑ Purchase order #________attached
❑ Charge my credit card (4 copy minimum):
❑ VISA ❑ Mastercard ❑ Access ❑ Eurocard
❑ Please Bill Me (4 copy minimum)

Card # ____________ Exp. Date ____________

Signature ____________

Name ____________ School ____________

Address ____________

City ____________ State ________ Zip ________

For more information, call Ms. Davis at 1-212-754-0539.
Expected shipment date Aug. 29, 1989.
Cut and mail this coupon to: 415 Madison Avenue, Dept. MPE, NY NY 10017
*Outside the U.S. add $1.00 per copy.

P9AA014

SCIENTIFIC AMERICAN

CORRESPONDENCE

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

Photocopying rights are hereby granted by Scientific American, Inc., to libraries and others registered with the Copyright Clearance Center (CCC) to photocopy articles in this issue of SCIENTIFIC AMERICAN for the flat fee of $1.50 per copy of each article or any part thereof. Such clearance does not extend to the photocopying of articles for promotion or other commercial purposes. Correspondence and payment should be addressed to Copyright Clearance Center, Inc., 21 Congress Street, Salem, Mass. 01970. Specify CCC Reference Number ISSN 0036-8733/89. $1.50 + 0.00.

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

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

Address subscription correspondence to Subscription Manager, SCIENTIFIC AMERICAN, P.O. Box 3187, Harlan, IA. 51593. Telephone inquiries: 1-800-333-1199, U.S. only; other 515-247-7631/32. The date of the last issue on subscriptions appears on each month's mailing label. For change of address notify us at least four weeks in advance. Please send your old address (if convenient, a mailing label of a recent issue) as well as the new one.

Name ______

New Address

Street ______

City ______

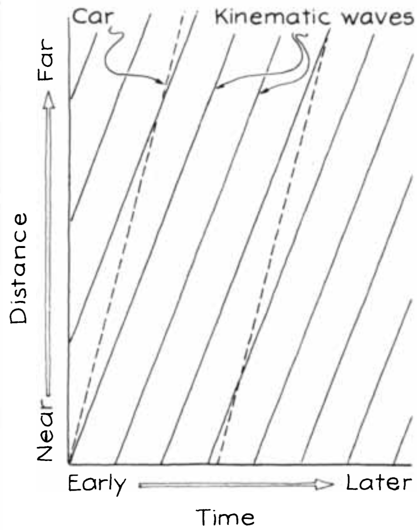
State and ZIP ______

Old Address

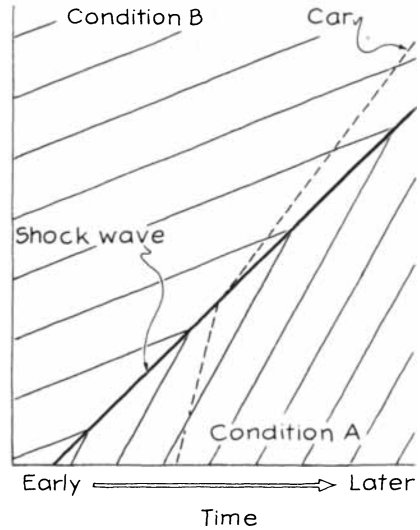
Street ______

City ______

State and ZIP ______



A distance graph



Graphing a downstream shock wave

through such a shock wave. Initially you are traveling at some modest speed in moderately concentrated traffic. Then you suddenly realize that the car ahead of you has slowed abruptly, and you hit the brakes to avoid a collision. Thereafter, you travel slower and in traffic with greater density than before. If it was a slow truck that produced the shock wave and if you are well behind it, the whole affair will be puzzling because you do not pass anything that accounts for the drama or the delay. If the truck gradually picks up speed, the traffic near it also picks up speed and thins out. The changes in speed and concentration travel back through the pack of cars, and eventually, perhaps much later, they reach you.

The shock wave I have described travels downstream because the line connecting *A* and *B* on the flow graph has a positive slope. You might investigate situations in which the line connecting two points on the flow graph has a negative slope, in which case the shock wave moves upstream or has no slope (the shock wave is stationary with respect to the road).

I decided to study the traffic flow on a section of expressway near Cleveland State University in order to see how it changes from midafternoon to the time when rush-hour traffic partially clogs the three outbound lanes. From a bridge spanning the highway I aimed a video camera to look upstream at the approaching traffic, and I ran tape for two hours. The camera had a convenient digital-clock feature that superposed the hour and minute on the recorded image of the highway. (It is important to mount the camera over the middle of the traffic flow being recorded; if the view is from one side, a large truck or bus in one lane can hide small cars in a farther lane.

Later, at home, I studied the tape and followed Lighthill and Whitham's recipe for constructing a flow graph. To establish an arbitrary distance *L*, I affixed narrow strips of masking tape across my monitor, one near the top and one near the bottom. To establish a time period *T*, I advanced the tape to a region of interest and let it run until the superposed minute symbol changed. That change marked the beginning of *T*, and the next change in the minute figure marked its end; during the 60-second interval many vehicles moved from the top strip to the bottom one on the monitor. I rewound the tape to the beginning of the interval, and when the first vehicle appeared I measured its strip-to-strip transit time with a stopwatch. Then I backed up the tape to time the next vehicle, and so on. When the minute symbol finally changed at the end of the measuring interval, the vehicle I was timing just then was the last to be included in the data for that interval.

To measure *L*, I could have noted particular features of the highway near the strips of tape on the monitor (skid marks, litter and so forth) and then returned to the scene to pace off the distance, but I chose a safer technique. When a school bus appeared on the videotape, I froze the frame and with a ruler measured the length of the bus on the monitor and also the distance *L*. By later measuring the actual length of a bus at a local school, I converted the measurements from the monitor to get *L*.

I took measurements for only the middle lane of traffic. That created a complication. When a vehicle changes lanes somewhere along L, should it be included in the data or not? If the flow in adjacent lanes is much different from that in the middle lane, the progress of such a vehicle will be different than if it had not changed lanes. I decided to include the data if the car was in the middle lane for at least half the length of L.

After taking data a number of times in the course of the two-hour videotape, I plotted a flow graph. Although the data points are scattered, a curve drawn through them roughly resembles the idealized flow graph of Lighthill and Whitham. The points reflecting the traffic flow before rush hour fall on the left side of the graph, and those associated with rush-hour traffic fall on the right side. One point high on the graph is about midway along the concentration axis and apparently represents the flow when it was near its maximum.

I spent additional hours watching the videotape without taking measurements (much to the amusement of my family). Although I saw cars speed up and slow down, I failed to detect distinct kinematic waves. They are just too subtle. I did easily spot shock waves, however. Before rush hour they usually moved downstream, although they sometimes moved upstream when for a moment the concentration increased markedly. In rush hour they always moved upstream. Every time a shock wave appeared—and particularly when it was accompanied by the sound of squealing tires—I was grateful that I had only to watch the wave, not to drive through it.

You might extend my analysis by studying the "traffic hump" that is created when there is a large influx of cars onto a highway where the flow has been moderate—say after a sporting event lets out and the fans head for home. Or you might dissect the flow associated with a "bottleneck," where a specific obstacle, such as an accident, blocks one lane or more. A hump can be short-lived, but a bottleneck can sometimes constrain traffic flow for an hour or more after the obstacle is removed.

FURTHER READING

ON KINEMATIC WAVES: II. A THEORY OF TRAFFIC FLOW ON LONG CROWDED ROADS. M. J. Lighthill and G. B. Whitham in *Proceedings of the Royal Society of London*, Series A: *Mathematical and Physical Sciences*, Vol. 229, No. 1178, pages 317-345; May 10, 1955.

COMPUTER RECREATIONS

A cellular universe of debris, droplets, defects and demons

by A. K. Dewdney

"Cellular automata are stylized, synthetic universes.... They have their own kind of matter which whirls around in a space and a time of their own."

—Tommaso Toffoli
and Norman Margolus,
Cellular Automata Machines

Certain mathematical systems, although easily described, can generate miniature universes of incredible complexity. These universes are not accessible by telescopes or spaceships; they can only be explored by means of computers. The Mandelbrot set (which I have described in this department on several occasions) is one example. Its members are found by iterating simple arithmetic operations, and yet, when displayed on the computer screen, they produce images of breathtaking intricacy. Cellular automata offer perhaps even better examples because, like our everyday universe, they change over time.

A new type of cellular automaton, discovered by David Griffeath of the University of Wisconsin at Madison, provides the best example yet of such a miniature universe. Started in a random state, it exhibits four distinct phases, ending with strange crystalline growths that remind one strongly of primitive forms of life.

A cellular automaton consists of an infinite grid of cells, each of which is in one of a number of possible states. Each cell changes its state in synchrony with the tick of an imaginary clock according to a set of simple rules. As implemented on a computer, the cells are pixels on the screen whose states are represented by different colors. Given the right set of rules and initial states, computer-based cellular automata can generate extraordinary patterns of color that evolve over time.

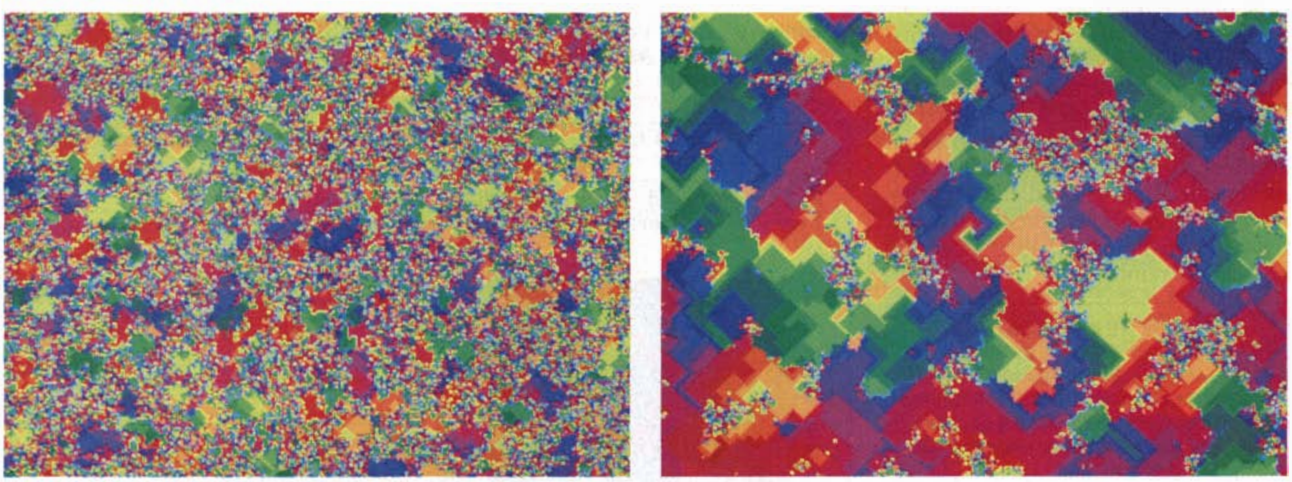
Griffeath's creation, which I shall call cyclic space, relies on an absurdly simple rule to produce striking phenomena of scientific interest and great beauty. The rule is based on the numbering of n states from 0 to $n-1$. A cell that happens to be in state k at one tick of the clock must "eat" any adjacent cells by the next tick that are in state $k-1$. Ingestion is indicated by a change of state of the adjacent cell, from $k-1$ to k. The rule resembles a natural food chain: a cell in state 2 can eat a cell in state 1 even as the latter is eating a cell in state 0. In cyclic space, however, the food chain has no end, because a cell in state 0 eats neighboring cells that are in state $n-1$.

On the basis of that one simple rule, cyclic space can turn a random distribution of colors into stable, angular spirals. Griffeath refers to the initial disordered stage as debris—and with good reason: the cyclic space looks cluttered at that stage. After some time, however, little droplets—awash with internal waves of color—succeed the debris. The droplets enlarge until they fill the cyclic space completely. At that point, crystalline spirals begin to develop. They are elegant growths that feed as they revolve.

Each spiral arises from what Griffeath calls a defect, a term borrowed from crystallography. The spirals grow in slow majesty, but they eventually compete for space. Some are overrun, whereas others remain to dominate the last phase of cyclic space. Griffeath calls the survivors demons.

Regardless of whether the number of possible cell states is small or large, the same scenario unfolds: debris is followed by droplets, defects and demons. The first three phases have the property of being metastable: they persist for a great many cycles before giving way to the next phase.

Consider a cyclic space that has 20 possible cell states. Why would an initial, random distribution of states be metastable? In other words, why does it take a long time for large areas of the screen to change their pepper-and-salt appearance? (To be sure, a



A 14-state cyclic space in the debris* (left) *and droplet* (right) *phases

few insignificant droplets pulse away here and there at the start, but they seem to be unimportant.)

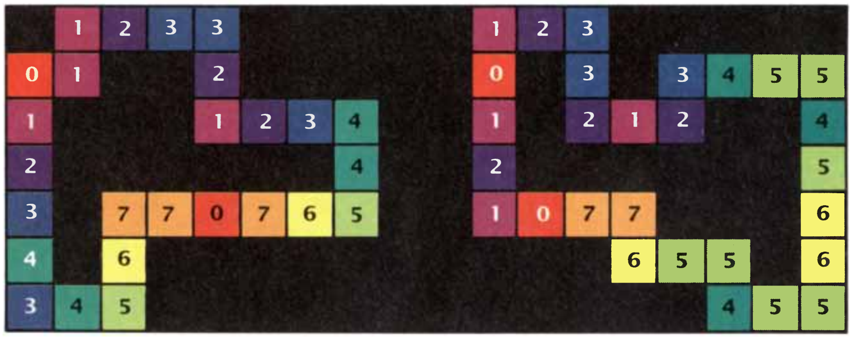
The answer is actually quite simple. If a cell happens to be in, say, state 5, what are the chances that at least one of its four neighbors will be in state 4? Not good, obviously. In the case of 20 states the probability is about .19. Even if one of the cell's neighbors does happen to be viable "food," what are the chances that either the original cell or the ingested neighbor will be next to a cell in state 4 at the following tick of the clock? This probabilistic argument explains what is seen on the computer screen when one runs a cyclic space seeded with random cell states: after a few sporadic meals, the cyclic space becomes stagnant.

Not entirely, however. What about those few isolated droplets? Waves of eating sweep across the face of each droplet, rebounding from the surrounding debris in new directions. That activity appears not only to sustain a droplet but actually to make it grow: new cells are recruited from the outlying debris to join the activity inside the droplet. The enlarging droplets gradually fuse together, leaving only a few scattered islands of debris. What causes the droplets to grow?

One guarantee of a droplet's continued enlargement is the presence of a loop: a closed chain of cells in which the state of each cell differs from that of its neighbors by no more than 1 [*see top illustration on this page*]. Hence, if one picks any cell in a loop, the next cell's state number is given merely by either adding 1, −1 or 0 to the picked cell's state number. (Remember that in cyclic space the highest state differs from state 0 by 1.) If one keeps a running sum of such differences as one goes from cell to cell in a complete circuit of the loop, it may happen that the final sum is not 0. A loop having that characteristic will sustain a droplet's growth indefinitely. It is in fact a defect, the very foundation of the third metastable phase of cyclic space.

The cellular action generated by a defect differs from that generated by a nondefect loop. A defect entrains surrounding cells to change states in a regular rhythm. On a computer screen the area surrounding a defect takes on an attractive spiral geometry. Eventually the entire cyclic space (or at least the part of it one sees on the screen) becomes completely covered by such spirals.

How do defects arise as cyclic space evolves from debris? Naturally, a defect might be present from the very beginning of the cellular universe. Indeed, if one considers cyclic space to be infinite in extent, then it is statistically guaranteed to have defects among its debris. Yet defects typically are formed where there were none before. How? Readers are hereby challenged to produce a small rectangle of cells that does not contain a defect but that will in time develop one. This means developing some simple test (short of letting the computer run for a long time) that detects areas where defects will develop.



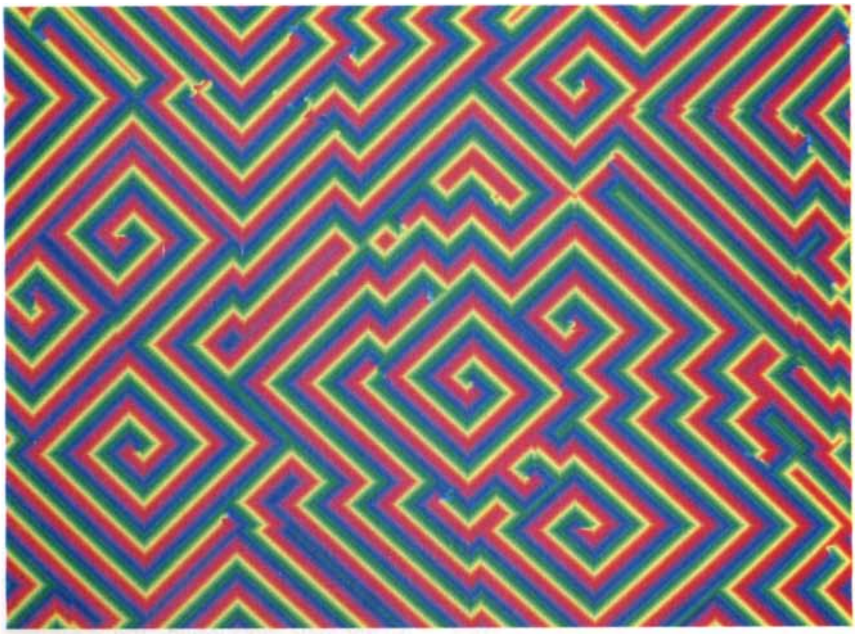
Two loops, one of which (right) *is a defect*

The fourth and final phase is hard to tell visually from the third. According to Griffeath, some defects are more "efficient" than others in that they function as *n*-state clocks: the march of cell states around the innermost loop takes exactly *n* ticks of the clock. Spirals that march to a slower beat end up being absorbed by their more efficient cohorts. Such efficient spirals are the demons that populate the fourth phase of cyclic space. They settle into an indefinite existence, marked off from neighboring demons by invisible but rigid fences.

Cyclic space might not have been discovered by Griffeath had not Norman Margolus and Tommaso Toffoli of the Massachusetts Institute of Technology developed the Cellular Automaton Machine, or CAM. The CAM consists of several large chips on a board that fits inside a personal computer. Within the chips each cell is represented by a memory location that is updated by a special control logic. At full speed the CAM can recompute a 256-by-256 cellular array at a rate of 60 times a second.

Unfortunately, such a device is rather expensive. Nevertheless, readers can write their own cyclic-space simulator (albeit one slower than a CAM) called DEMON by translating the following algorithmic outline into their favorite programming language. The algorithm uses two arrays called *new*


The cyclic space in its defect phase



The cyclic space in its final, demon phase

and *old* to hold the current and previous states of the cells in cyclic space.

```
repeat until key is pressed
for i ← 1 to 100
    for j ← 1 to 100
        for each neighbor (k,l) of (i,j)
            if old(k,l) = old(i,j) + 1
                then new(i,j) ← old(k,l)
for i ← 1 to 100
    for j ← 1 to 100
        display new(i,j)
        old(i,j) ← new(i,j)
```

The outer loop, which ends when the user of DEMON presses a certain key, controls repetition of the main checking and displaying cycles. (Of course, there are other ways to construct such a loop.) Readers who have computers with small memories are advised to limit themselves to a smaller array of cells, say one that is 50 by 50. The two inner loops used here presuppose a 100-by-100 cellular array. The cell (i,j) sits at the intersection of the ith row and jth column.

The innermost loop of DEMON simply checks the state of each of the four cells adjacent to cell (i,j). (Readers must therefore include instructions in their versions of the algorithm that assign the values $i-1$ and $i+1$ to the index k while l equals j. Similarly, the values $j-1$ and $j+1$ must be assigned to l while k equals i.) If the state number of a neighboring cell happens to be 1 more than the state number of the cell at (i,j), then the cell at (i,j) is eaten: it has its state number changed to that of the neighboring cell.

DEMON must perform modular arithmetic when it calculates the value of $old(i,j)+1$. In other words, if $old(i,j)$ happens to be $n-1$, the highest state number, then $old(i,j)+1$ is equal to 0. Hence, if one has specified, say, 10 states in the cellular space, the state numbers will be 0, 1, 2..., 9 and $9+1=0$. If the value of a cell changes, the new value is assigned to $new(i,j)$.

The double loop that follows the checking loop displays all the cells in *new* and then updates the *old* array by replacing all its values by the corresponding values in *new*. Readers can consult previous columns, including last month's, to figure out how to display the cells on the computer screen.

Modular arithmetic must be applied not only with cell-state numbers but also with the indexes i and j themselves. The simplest way to achieve the illusion of an infinite space is to endow one's screen with the "wraparound" property. Cells on the extreme right-hand margin are considered to be adjacent to those on the left, and cells at the bottom of the display are effectively adjacent to those at the top. The effect is created by using index values from 0 to 99 instead of 1 to 100. The cell to the immediate right of (23,99) is in fact (23,0). Hence, the numbers $i-1$, $i+1$, $j-1$ and $j+1$, which are index values of the neighbors of cell (i,j), must all be expressed in modular form. Most programming languages have instructions that do that automatically.

Naturally, DEMON must allow a user to "initialize" the cellular space under its control. This step can be done by including a provision in the algorithm for eliciting the desired number of states from the user, as well as a loop that gives every cell in the space an initial, randomly chosen state within the allowed range of numbers.

What number of states works best? It all depends on how long the reader is willing to wait for the four phases to succeed one another. When the number of states is very large, say more than 25, a 100-by-100 random cellular array is likely to remain fixed forever. On the other hand, when the number of states is small, the stages succeed each other too quickly to be appreciated. Griffeath recommends between 12 and 16 states.

Some months ago I mentioned that I would visit the University of Wisconsin to report on the activities of the so-called particle mafia, of which Griffeath is a member. (The name comes from an early association of the general field of particle systems with Frank Spitzer, a mathematician at Cornell University in upstate New York, where—according to Hollywood legend—the Mafia used to hang out.) Spitzer's articles and talks on particle systems first popularized the subject in North America, just as the work of his colleague R. L. Dobrushin did in the Soviet Union. Several of Spitzer's students at Cornell as well as his followers elsewhere took up the work of controlling the particle-systems numbers game. They include not only Griffeath but also Maury Bramson (also at Madison), Richard Durrett of Cornell and Thomas Liggett of the University of California at Los Angeles.

What exactly is a particle system? It is usually a cellular automaton in which only one cell changes at a time, often in a random fashion. The diffusion-limited aggregation algorithm described in last December's column is an example of a particle system. There, a single particle wanders randomly across a grid of cells until it encounters a growing aggregate. Its position is then frozen as another particle begins to wander. In time, the aggregate invariably develops a treelike shape.

A number of important but difficult problems are raised by particle systems, not the least of which are questions concerning the long-term stability of certain phases that arise when the systems are set running. Although cyclic space is not a traditional particle system, Griffeath thinks it nonetheless provides a model for locally periodic wave formations that, because of its straightforward rule, is particularly amenable to mathematical study. Indeed, he has recently proved a theorem that gives, for each possible array size and number of states, an estimate of the probability that cyclic space will never escape from debris gridlock.

For the rest of us, the cyclic cellular automaton at least provides an opportunity to explore a miniature universe. To be sure, it is much simpler than the universe in which we exist, but it has alien beauties and wonders of its own.

Last April's foolery on the subject of matter fabrication brought a gratifying response from the great majority of readers who recognized that "Arlo Lipof" (the name given to the person highlighted in the column) is an anagram of "April Fool." The column featured two methods of matter fabrication, both based on paradoxes. The first method involved cutting up an 8-by-8-inch square into four pieces and then reassembling them to obtain a 5-by-13-inch rectangle. If the operation is applied to a square gold bar, the excess gold amounts to an immediate increase in wealth.

Most readers solved the "paradox" by examining the diagram of cuts very closely. The reassembly shown in the illustration is faked (as Lipof clearly stated). If one cuts out the requisite pieces and attempts to reassemble the rectangle, a narrow gap shows up in the middle of the figure. Its area is exactly one square inch. Hence, the first paradox is no paradox at all but an old chestnut that I first discovered in the charming book *Tangrams—330 Puzzles,* by Ronald C. Read.

The second paradox is perfectly real even if, contrary to Lipof's claims, it cannot be applied to matter fabrication. The Banach-Tarski paradox states that it is possible to cut a solid sphere into a finite number of pieces in such a way that the pieces can be reassembled into an equally solid sphere of twice the volume! The exclamation mark may not seem fully deserved to those readers who realize that the "pieces" are not at all simple and are best described as mathematical sets.

The illustration accompanying my description of the paradox was adapted from the book *The Banach-Tarski Paradox,* by Stan Wagon of Smith College. In the illustration, two-dimensional hyperbolic space is projected in two different ways onto a flat disk. It provides a visual demonstration of the paradoxical nature of the space. Three sets are shown, any one of which is simultaneously equal to one half and one third of the space! The theorem underlying the illustration is the result of the joint work of Wagon and Jan Mycielski of the University of Colorado at Boulder.

The Banach-Tarski paradox actually made headlines recently in mathematical circles, since it is related to the work of Miklós Laczkovich of Loránd Eötvös University in Budapest [see "Science and the Citizen"; SCIENTIFIC AMERICAN, July]. Laczkovich has proved that a circular disk can be decomposed into finitely many pieces (fewer than 10^{50}, in fact) that can then be reassembled into a square with no space left over.

Readers of Dennis Shasha's book *The Puzzling Adventures of Dr. Ecco* were challenged to earn Ecco's title of "omniheurist" by solving 10 encrypted puzzles. The omniheurist contest winners are Fred Galvin of Lawrence, Kan., and Lambert Bright of Lincoln, Neb., who also won the grand prize of a chess set awarded by lottery.

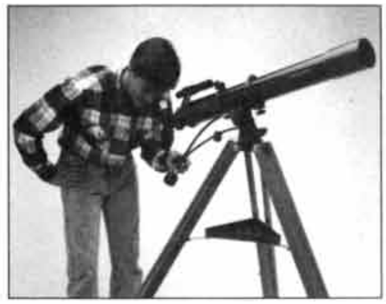
One of the puzzles in the omniheurist contest called for designing a layout for an Antarctic research building. The building had to have 31 rooms, all with 20 square feet of floor space. No room could have more than four doors, and wall thickness was assumed to be negligible. The 31 rooms and their doors had to be arranged so that one could walk from any room to any other room by passing through no more than six doors.

The best solution to this puzzle was sent to Shasha by Brian L. Platt of Woods Cross, Utah. Readers can try to duplicate Platt's success by thinking of a tree structure. It may take an actual layout to convince oneself that a reasonable architecture is possible. Omniarchitects might discover a more advanced design in which more than 31 rooms can be traversed under the same constraints. Candidate plans should be mailed to Shasha at the Department of Computer Science, Courant Institute of Mathematical Sciences, New York University, 251 Mercer St., New York, N.Y. 10012. I shall publish the name of the winner in a forthcoming column.

FURTHER READING

CELLULAR AUTOMATA MACHINES. Tommaso Toffoli and Norman Margolus. The MIT Press, 1988.

CYCLIC CELLULAR AUTOMATA IN TWO DIMENSIONS. Robert Fisch, Janko Gravner and David Griffeath. Ted E. Harris Festschrift. Birkhäuser, in press.




THE POWER OF
ORIGINAL THINKING.

BOOKS

Where "three" began, wind and fire, a false dichotomy, C. V. Raman's effect

by Philip Morrison

In Search of the Indo-Europeans: Language, Archaeology and Myth, by J. P. Mallory. Thames and Hudson, 1989 ($29.95).

The swastika with its broken symmetry, the fourfold rotations without reflections, is a widespread primeval design. In our times it became uniquely "the symbol of a terror more terrible than the snake-girdled Medusa's head." (So Hermann Weyl said to a divided mathematical audience in Vienna in 1937, and his words were answered by pandemonium.) The hooked cross still evokes the foreboding of evil even when you encounter it carved long ago in all its enigmatic innocence on the frieze of some South Indian temple.

The grotesque obsession with the Aryans of the swastika (otherwise the Indo-Europeans) developed after 1870 from what is actually a genuine linguistic heritage of some two billion people who "cannot count . . . or express our most fundamental actions . . . without recourse to a . . . system of speech" shared for 6,000 years. That fact has nothing whatever to do with the horrid myth of Aryan supremacy and racial purity, now justly anathema. Indeed, Yiddish is as Indo-European as any other German dialect, and the strongest claim in northern Europe to the label of Aryan is that of the Romany-speaking Gypsies.

In this learned but fast-flowing volume an archaeologist at Queen's University in Belfast lets the reader in on a contemporary domain of research that each year blossoms forth in some 2,000 papers: philology and linguistics, archaeology, comparative mythology and religion. (He says nothing at all of any relevant biochemical genetics, still too contaminated by the toxic debris of the Third Reich.)

Every literate person knows at least vaguely of the Indo-European concept, but Professor Mallory cites the kind of evidence, fresh to a lay reader, that for two centuries has tantalized scholars of the word. At the simplest level you see a list of numerals in two dozen languages, modified from a similar list published by the pioneer, dilettante James Parsons, in his "gullible" and "tedious" volume of 1767. Try a sample of words for the numeral three in tongues whose speakers span half of Eurasia: Irish, Bengali, Lithuanian and Tocharian (an old language known in manuscript from the Tarim Basin in Chinese Turkestan). In that order they read *tri, tri, trys, tre.* (Controls are prudent: compare words for the third integer in four non-Indo-European tongues, Turkish, Hebrew, Malay and Chinese: *üc, səlosa, tiga, san.*)

Similarities simple and subtle have by now been elaborately mapped out for both sound and structure among some 50 modern languages and as many historic ones, back to the oldest witness. That is offered by cuneiform tablets of Assyrian merchants, excavated at the trading posts they established in Anatolia. The tablets record the Indo-European personal and place names of local trading partners; the clay documents go back to about the 19th century B.C., older than any Hittite, Sanskrit or Aegean texts.

The hypothesis that "cannot be avoided" is that somewhere in Europe or Asia there was a place where long ago people spoke the protolanguage—the common ancestral language. (Only the myth-dazzled will accept the proposal that the people lived at the North Pole, an inference from certain poems that can be read as fixing the North Star at the zenith of the old homeland.) The grand goal is to find this focus in space and time and to finger its rich human texture; the great problem is how to gain so concrete a result from evidence that is almost entirely linguistic. "No Proto-Indo-European text exists; . . . physical remains . . . cannot be identified without extensive argument; and their . . . location has been the subject of a century and a half of . . . debate."

Writing is older in Asia, and there we find the earliest clues. The inferred vocabulary of the Indo-European legacy surely includes domestic animals, agriculture and pottery, and so its time was surely Neolithic or later; indeed, the vocabulary seems to go beyond the dawn of that way of life on to the time of wheels, of dairy products, of ploughs, of the magnificence of war chariots and the horse. From the Danube to the Caspian we have found an intricate archaeology of the wheel, beginning with the steppe-adapted ox wagon. The breaking of horses too begins about that time. The community of early Indo-Europeans must, then, have formed its tongue and its terms by about 4500 B.C., a time set prudently a little before the oldest evidence at hand.

Where to insert the spade? First one needs to grasp the nature of the society; that will give scale and perhaps rough location. A linguistic argument is a surprising help: the speech of those dispersed faraway from a center tends to become conservative; that of people closer to the focus is refreshed and changes more readily with the times. (Quebec French and Appalachian English echo the Paris or London of centuries ago.) That clue points to a center that is not at the geographic fringe, neither west to Galway nor east to the Tarim, but central. Linguistic communities in the relevant epoch seem usually to occupy an area of from a quarter of a million to a million square kilometers. We are looking for the relics of such a culture.

Somewhere between the Black Sea and the Volga-Ural plains is the consensus choice. The Soviet archaeologists have found candidates that begin to fit. A thousand sites of a village culture of about the right time, or a little early, dot the Ukraine; even closer fits are abundant farther to the east. But the match is uncertain. Recent reconstruction of Indo-European religious ritual and class organization as a tripartite society of priests, warriors and herder-cultivators has much impressed the scholars. It is not easy to read that scheme, however well it may fit later practices and texts, into the preliterate copper rings and ceramic shards, the wheel planks and horse trappings that are the grave goods of the steppe.

Up-to-date revisionists of time and space are at work, but Professor Mallory argues explicitly and plausibly for his mainstream view. What is missing, what will bring the result so long sought, will be good evidence of the expansion of one of the regional cultures recognized under the spade. That will not be any conquest by blue-eyed heroes; rather, we seek a subtle spread of new opportunity and societal welcome that could bring to many neighboring regions a bilingual phase, with the new speech winning in the end. The issue is open; the matter may now be far enough beyond the murderous fervor of its abuse to rejoin the list of high puzzles within the science of prehistory.

WIND AND SEA: STATE OF SEA PHOTOGRAPHS FOR THE BEAUFORT WIND SCALE, by W. T. R. Allen. Environment Canada, Canadian Government Publishing Centre, Ottawa, 1983 (paperbound, $7.95 in Canada, $9.95 in other countries). **THE GREAT BLACK DRAGON FIRE: A CHINESE INFERNO,** by Harrison E. Salisbury. Little, Brown & Company, 1989 ($17.95).

Captain Francis Beaufort (in due time he made admiral), oceanographer to the Royal Navy at the time of Nelson and on through the *Beagle* decades, carefully classified wind and waves at sea into 12 grades, using sea state to fix wind speed. His laconic descriptions are as eloquent as they were practical, evoking the mariner's fear and delight as well as signaling when it became prudent to furl your lofty toproyals. (This bilingual text cites only more modern and somewhat less vivid versions.) In this set of photographs made afloat, landlubbers can see those well-assessed environments redefined in modern color images, framed by the steel decks, railings and rigging of a seakeeping Canadian weather ship. The seascape is the open Pacific at Station PAPA, 1,000 lonely miles seaward from Vancouver.

At Force 0, extremely rare out at Station PAPA, the sea is a mirror, although some swell passes smoothly beneath the still air. Moved by the fresh breeze of Force 5, moderate wave crests six feet high begin to lengthen, and white caps are many. Force 11 was encountered only once during the five-year study. Exceptionally high waves up to 52 feet (about one wave in 10 reaches maximum height) move in the violent storm, their towering crests everywhere blown by 68-knot gusts into a froth that fills the air. Force 12, the hurricane itself, was never met by either of the two ships *Vancouver* or *Quadra* while they shared the station.

Two meteorologists at sea, R. F. Webber and J. H. Scarlett, took 338 pictures at PAPA between 1976 and 1981; most of the steady-state sea and wind states are illustrated by two photographs chosen for each. The intent is to aid a ship's officers in their wind estimates; the outcome is the visual extension of the reader's experience, a remarkable extension of old Beaufort's scientific seamanship. His descriptions were calibrated against knots of wind speed only within this century; they have long communicated aptly what is here plainly seen.

The element of fire has been restive in fact and in thought during the past few years. In the latter part of 1987 one of our best-known correspondents, Harrison E. Salisbury, went to the remote scene of a terrifying conflagration that had raged in May of that year along the Black Dragon River (we call it the Amur), which parts the northernmost region of China—once Manchuria—from Soviet Siberia. That "sea of green," the main Chinese share of the great conifer forest that covers northern Eurasia, burned furiously for almost a month, flames engulfing the lumber towns of the land, until the fire was controlled by rainfall—some of it following dry-ice cloud-seeding by artillery and aircraft—and by the backfires and unstinting effort of 40,000 firefighters, most of them army units long stationed in that border region.

The area of forest felled and blackened on the Chinese side of the river is about 5,000 square miles; the Chinese fought hard for lives, towns and timber. On the Soviet side the remoteness of the big forest made it possible essentially to ignore the fire, although the Siberian larches there burned over perhaps five times the area in China—more than all of Vermont and New Hampshire. This was the largest well-documented fire in centuries, an order of magnitude more extensive than the Yellowstone burn of 1988 or the Minnesota fire of 1918. (A strong case has been made that in 1915 the conifer forest burned over a wide swath in central Siberia between Lake Baikal and the Urals, felling an area that was an order of magnitude greater than even the Black Dragon fire.)

Fire began on the Soviet side, likely set by lightning; it was noticed in satellite images but was not made known to the foresters south of the border. A summer of drought and a winter of little snow had been followed by a spring of unusually strong winds, Force 6 to Force 9. The push to harvest timber had left a deep cover of waste on the ground over large areas. The fire became a media event in China, giving the nation one of its earliest TV celebrities: Black Beard, the tireless field commander of an army unit that performed with distinction in the firefighting campaign. (Against custom and regulations he did not shave for the duration.) That the fire was widespread in origin did not prevent the later trial and conviction of a hapless young culprit for starting one outbreak carelessly in a visible place.

Will the debris burn again, as many experts expect? Can 10 million tons of dead timber be removed along the single rail line before rotting? Was the forest itself the new-growth result of a fire long ago? Will the vanished wildlife (no tigers, but many bears) return? What will happen to the climate nearby? (In Beijing, 1,000 miles south, the streetwise expect dust storms.) These questions are open.

The author himself grew up in timber country and evokes with quick sympathy and insight the engaging people of this wide timberland: a decisive woman mayor, a thoughtful lifetime forester, truckdrivers and sawyers whose style in community dancehall or private comment is a Han echo of his own childhood milieu. The photographs are helpful; the book, lacking an index, is a brief, informed impression of an event that deserves attention worldwide. After all, at a minimum we people are united by a shared concentration of carbon dioxide.

PEACEMAKING AMONG PRIMATES, by Frans de Waal. Harvard University Press, 1989 ($29.95).

A primatologist at ape and monkey colonies from Arnhem to Madison and San Diego and a much praised animal photographer, the author has produced this vivid and original account at firsthand. His title has a hidden barb, for there is little doubt that he has wider goals, seeking to make some peace among human primates too—between observers in the field, who may come back with sparse data taken under pristine conditions, and zoo workers, whose data are copious but refer to animal societies under serious constraint; between strict geneticists of human behavior and subtle cultural anthropologists; even between NATO and the Warsaw Pact.

Nobody should claim that the truth is all on one side of what is often a "false dichotomy." Genes appear to hold most of destiny for the four simian species he pictures and describes. Yet de Waal, himself a Netherlander, reminds us that genes are not all. In Europe the people of that crowded nation hold "tolerance" as a key to their national character, whereas in South Africa their closest kin adhere tightly to a quite contrasting attitude. For de Waal, aggression is neither absent nor dominant; it is one fundamental trait in animals, along with which "powerful checks and balances . . . evolved to mitigate its effects."

Chimps, pygmy chimps (now called bonobos), rhesus and stump-tailed monkeys all look out at us here from fine portraits and scenes of intimate action, often X-ratable. The tightest arguments are based on depth of observation—six thousand hours watching those two dozen Arnhem chimps alone—but he spares the reader long tables of counts and probabilities, citing specific cases of what his data show, anecdote tempered by statistics.

Crowded rats kill each other. Monkeys show less dramatic effects. The Arnhem chimps, used to acres of free range, are crowded into one heated hall for the winter at 20 times their summer density. They groom and greet each other much more frequently than in the summer, actively trying to reduce social tensions. Yet they can kill too: he describes what seems to have been a joint fatal attack by two males on a third one who had newly challenged their long-shifting status as top-ranking male. But chimps also know how to make up: no flowers, but plenty of hugging, kissing and fondling, seen between males statistically twice as often as between females.

De Waal has seen a lot among rhesus monkeys, even a version of social class. A dozen animals divided themselves into two groups; within each group, the order of drinking during an experimentally contrived water shortage was "flexible and easygoing," and yet all members of the upper class took precedence over those of the lower. For other purposes the groups were not distinct. In the rhesus physical violence is not exceptional, as it is among most primates. Still, reconciliation after attack took place frequently, except that a friendly reunion was very unlikely after an attack by an upper-class female on a lower-class one.

The author's observations on the bonobo are unexpected. That species, neither pygmy nor chimpanzee (although at a glance most chimplike), has been distinguished from, and classified in one genus with, the more familiar chimpanzee only since 1929. Found only south of the big river in central Zaire, they mostly knucklewalk along the forest floor. The zoo in San Diego has nurtured a real colony, their first bonobo mother having borne 10 children within 14 years. In captivity they are tool users, show self-recognition in a mirror and are clearly of high social intelligence and good humor.

Bonobos are the sexiest of known primates. They couple face to face in short encounters—as juveniles, adult males with males and adult females with females, both in the wild and in the zoo. (Full acts of copulation with ejaculation are confined to adult male-female pairings.) This frequent sexual activity is not just for fun and not directed toward conception but is "an alternative" to hostility: making love not war. Most mounts and mating occur in tense conflict situations between partners. "Sexual conflict resolution is the key to bonobo social organization."

It is worse than fruitless to ask which primate species we most resemble. We share with all, and they with us. Our culture has plainly transcended theirs; warfare has gone past emotion, however dependent on it. Linnaeus said in later life that he had regretted forming the genus *Homo* just for us; today we might make a third species within the genus of chimps and bonobos, as *Pan sapiens.*

Yet it seems to a reader that the newly seen species-by-species diversity of primate behavior only strengthens Linnaeus' first decision: we are far too different, billionfold, with fire, language, domesticates, artifacts . . . , although by no means alien to our poor country cousins. Forgiveness is not a new Judeo-Christian idea, the triumph of reason and revelation, but rather a primate stratagem 30 million years old. It is time to call on it again, to call on all the mechanisms for turning from confrontation to negotiation, once more "adapting," as de Waal ends his book, "to new circumstances in our own self-interest."

Journey into Light: Life and Science of C. V. Raman, by G. Venkataraman. Indian Academy of Sciences, 1988; distributed in the U.S. by Oxford University Press ($45.00).

"Western science came to India before Western education." For a long time it was the exclusive territory of Europeans. By 1888, when C. V. Raman was born, there was a thin net of colleges and a few universities for the modern education of bright Indian students. That extraordinary youth became the first Asian Nobel laureate and a lifelong champion of independent research, of which he himself was a versatile and tireless practitioner. (A surprising number of talented scientists share descent from Raman's father, although not this biographer. South Indian names are personal, not regularly indicative of kinship; they are the names of deities.)

He was a brilliant student, gold medals in English and physics; he recalled being moved by the works of Buddha, by Euclid and by Helmholtz's text on sound. He took his M.A. in physics at the Presidency College in Madras under kindly Professor R. Llewellyn Jones, who left him strictly on his own. In the course of two years' work "I remember attending only one lecture, and that was on the Fabry-Perot." Yet somehow he published a small paper on oblique diffraction in *Phil. Mag.,* despite the fact that the journal was not known to the college or university library. A second paper on the perception of pitch drew a letter from Lord Rayleigh himself.

Plainly his heart was in science, where there were small prizes to be won (and he won them all) but no jobs. Not yet 20, he took the examination for the Financial Civil Service. "I took one look at all the candidates . . . , and I knew I was going to stand first." While waiting to be posted off, he married a young girl, Lokasundari, whose music had attracted him; it was not an arranged marriage, but it was a long-lived success. One son recounts that Lady Raman—a wry wit on her own—often said she did not know if Raman had married her for the extra allowance given married officers! He had, in fact, refused a dowry.

Then off to the capital, Calcutta, where the young couple set up house. The city was the center of science in India. By good chance Raman came across an idle and deserted laboratory, a dozen rooms built 16 years before for the Science Association, formed rather on the model of the Royal Institution in London. That body indeed sponsored a fine set of advanced lectures, but hopes that a research director might be found had come to nothing. Raman appeared as if from nowhere and offered to undertake research in physics without pay. His wife remembered their double life in those years: "5:30 A.M. Raman goes to the Association. Returns at 9:45 A.M., bathes, gulps his food . . . and leaves for his office. . . . At 5 P.M. Raman goes directly to the Association on his way back from work. Home at 9:30 or 10 P.M. Sundays, whole day at the Association." After a while they moved to a house adjoining the lab, and he had a door cut between home and lab. Meanwhile, of course, he was a full-time civil servant, praised by his chief as "one of our best men."

Raman always spoke of Calcutta as his Golden Era. He and his loyal assistant Ashu Babu worked to produce 27 publications, largely on sound, in 10 years. For instance, a clever photographic study showed that a transversely vibrating string moved enough horizontally at its nodes to transfer energy through them. After that decade he had attracted enough attention to leave the civil service for a professorship at Calcutta University, at half his government salary. He had become a professional physicist: an intense idealist, successful and deeply original, especially in the physics of waves, to which he applied his own ingenious touch.

Raman found his wavelength-shifted scattering of light in the late 1920's—a wholly quantal effect—and a Nobel prize came for it in 1930. Then he broke with Calcutta in a controversy with the theorist Meghnad N. Saha, one of the other great men in physics of that place and time. One virtue of this admiring but by no means wide-eyed biography is its excellent explanations of the physics of Raman's work, of many properties of waves in light and sound, of the Raman effect itself and of the remarkable later work involving the diffraction of light from a pattern of ultrasound in fluid.

So far that is only half a life, and less than half of a book. That beginning demonstrates through the singular strength of one man how it is that the river of science can flow through every land down to a single sea. Raman was no Englishman, however deft his English rhetoric, however meticulous his papers in *Proc. Roy. Soc.* or *Phil. Mag.* He was a South Indian, who lived in his own land mostly under foreign rule, a nationalist by poetry and music and dress. He was a natural philosopher, with uncharacteristically little interest in religion or its practices but well armed with the tools of 300 years of modern science. He celebrated the wealth of his culture, the excellence of mind and hand, the beauty of the natural world. He endured controversy; he was wrong on some issues, and he came to know it. The strong ego grew lonelier and finally, stubbornly, a little eccentric.

He was cremated in 1970 at a place they marked with a small tree, which is flowering now for the first time, in the quiet gardens of the vigorous Raman Research Institute he caused to be built once he was done with larger responsibilities. The course of his life parallels the course of modern India and its heroes. His India still confronts great issues of self-reliance, devotion and integrity, well put here in the context of science and technology by the thoughtful biographer, but nowhere in the world within easy reach of solution.

ESSAY

Fetal research: the underlying issue

by Robert J. Levine

Fetal research has led to the development of a vaccine for polio, to important fundamental insights into the biology of cancer cells and to improved treatments for serious diseases of both the fetus and the expectant mother. Now it seems likely that transplants of human fetal cells will ameliorate such devastating maladies as Parkinson's disease and juvenile diabetes. Even this incomplete list should serve to demonstrate the enormous value of fetal research. Moreover, there is by definition no other way to secure the benefits of such research, since federal law allows it only when a particular benefit cannot be realized by other means. The opponents of fetal research are demanding, in other words, that society relinquish all claim to such benefits.

These opponents are politically powerful and persistent, and they have been intermittently successful. Those who have been most vocal of late have as their ultimate goal the interdiction of elective abortion, and they see the outlawing of fetal research as a means to that end.

In the past 15 years many bills have been introduced in the U.S. Congress and in various state legislatures to ban such research; some of the latter have succeeded. And on several occasions moratoriums have been imposed on the use of federal funds to conduct some kinds of fetal research. The first moratorium was imposed by the National Research Act of 1974. It halted all "research ... on a living human fetus, before or after the induced abortion of such fetus, unless such research is done for the purposes of assuring the survival of such fetus." The act set up a national commission charged with recommending the circumstances, if any, under which the moratorium should be lifted.

After careful study of the relevant medical, scientific, legal and ethical issues, the commission recommended that the moratorium be lifted, and it was. The commission's further recommendations laid the basis for current federal regulations protecting the fetus as a research subject. The regulations provide elaborate and stringent safeguards for living fetuses, but all they have to say about research on the dead fetus or its cells, tissues or organs is that it must comply with applicable state or local laws.

The most recent federal moratorium, promulgated in May, 1988, by the then assistant secretary of health, was prompted by a proposal to use Public Health Service funds for research on the implantation of human fetal brain cells to treat Parkinson's disease. The moratorium forbids federal funding of research "utilizing human fetal tissue, obtained from induced abortions, for therapeutic transplantations." Again an advisory panel was convened to study the issues and make recommendations. The panel reported to a committee advising the director of the National Institutes of Health, and the advisory committee thereupon voted unanimously to end the ban. The office of the Secretary of Health and Human Services received the committee report in January but has not yet responded to it, and the moratorium remains in effect.

Note that this moratorium deals only with the use of tissues and cells of dead fetuses. The 1974 commission had considered such use to be relatively unproblematic; its central concerns were, rather, over the possibility that a living fetus might be harmed or wronged to serve research interests. Although the morality of abortion has always figured in the debate over the ethical permissibility of fetal research, it is notable that those who oppose the transplantation of fetal tissues seem to consider it the only issue.

Their attack on fetal research is grounded partly on the incorrect belief that doctors encourage abortions in order to get research material. Investigators, who do not themselves perform the abortions, are portrayed as accomplices whose work creates the impression that the "abortion industry" is legitimate. Fetal research is also denounced as being unethical because the subject cannot give informed consent and no one appears to have the standing to give proxy consent—surely a mother who has chosen to end the life of her "unborn child" cannot be relied on to guard its interests. If research is permitted on those who cannot consent, it is asked, who will be next? The mentally retarded? Social undesirables?

It is further alleged that a woman who for moral reasons is ambivalent about having an abortion might decide to proceed if she is reassured that some good might come of it—for example, that a Parkinson's-disease patient might receive therapy. Worse, a woman might choose to get pregnant merely to produce research material; perhaps she could sell her fetus or give it to a relative who has diabetes.

Now I have provided a partial list of objections to fetal research to balance the partial list of benefits I mentioned at the outset. Having considered these and other benefits and objections, I believe that work in the field of fetal research should continue and that the federal government should support it.

I further believe that investigators should be respectful of and responsive to the concerns of those who oppose fetal research. Actually, most of these concerns are already accommodated in federal regulations. Among other things, the regulations forbid financial or other "undue" inducements and require separation of a woman's decision to have an abortion from her decision to permit research. All proposals to conduct fetal research must be reviewed and approved in advance by interdisciplinary committees. The regulations also provide ample barriers to the abuse or exploitation of the mentally retarded and other vulnerable people.

One issue that cannot be resolved by writing federal regulations is that of the morality of abortion. Given the nature of the controversy, it seems highly unlikely that this issue can be resolved at all.

It is not necessary, however, to justify induced abortion in order to justify fetal research. Abortion is not now and never should become a means to achieve investigative goals. Many women will choose to have abortions whether or not there is fetal research. Given that reality, the question is what to do with the remains of the fetus. It can be put in the incinerator along with impersonal surgical specimens. Or it can serve as a tissue donor for research or transplant purposes in accordance with the provisions of the Uniform Anatomical Gift Act.

Surely the latter course of action is the more responsive to whatever claims for human regard the dead fetus may have.

ROBERT J. LEVINE is professor of medicine and chairman of the institutional review board at the Yale University School of Medicine.

國際
胡金銓
編導
鵬楓鷹雋
Showdown in Taipei.
Today's corporate warrior in Taiwan needs more than his own skills. He needs a fleet at his command.
The fleet of United nonstops to Taipei.
With more nonstops to Asia and the Pacific than any other airline, United lands you there ready for action. And each flight provides the best in international travel. Including, in First Class, sleeper seats and our exclusive Concierge Service.
United. Rededicated to giving you the service you deserve. Come fly the friendly skies.
UNITED
AIRLINES
TOKYO • OSAKA • HONG KONG • SEOUL • TAIPEI • SYDNEY • BEIJING • SHANGHAI • AUCKLAND • SINGAPORE • BANGKOK