SCIENTIFIC AMERICAN

MEMORY IN BIRDS

$2.50

March 1983

*Sedan 35 est. hwy. 27 EPA est. mpg. Town & Country Wagon 32 est. hwy. 24 EPA est. mpg. Convertible 29 est. hwy. 22 EPA est. mpg. Use EPA est. mpg for comparison. Your mileage may vary depending on speed, trip length

LeBaron Coupe is equipped with these standard luxury features:

- Power front disc brakes •
- Power rack-and-pinion steering •
- 5-speed manual transmission •
- 2.2 liter engine •
- Low-back bucket seats with console •
- New instrument panel message center •
- Padded landau vinyl roof •
- Halogen headlamps •
- Maintenance-free battery •
- Electronic digital clock •
- Luxury steering wheel •
- Deluxe wheel covers •
- 20 oz. cut pile carpeting •
- Dual manual mirrors •
- Warning chimes •
- Crystalline pentastar hood ornament •

Mark Cross Package with leather interior, optional.

& weather. Hwy. & CA mileage lower. **Base sticker price excludes title, taxes, destination charges. †Whichever comes first. Limited warranties. Deductible may be required. Buy or lease at your Chrysler-Plymouth dealer.

They are like no other cars in America, Europe or Japan. And Chrysler has the confidence to back each car with a 5 year or 50,000 mile protection plan.

Chrysler brings back the romance of the past with the luxury and technology of the present: the Town & Country Convertible. It's like no other car on the road with classic "woody" styling, a Mark Cross leather interior, power top and advanced electronic instrumentation.

Each LeBaron is a unique statement of luxury. There's personal luxury in the coupe. The luxury of six passenger room in the sedan. Functional luxury with the Town & Country wagon. And the ultimate luxury in a choice of unique convertibles.

And every LeBaron is pure driving luxury. With front-wheel drive and power steering, you enjoy remarkable control and a smooth road-hugging ride.

A Quality Assurance Program assures quality so exceptional that Chrysler has the confidence to protect LeBaron's engine, power train and its entire outer body against rust-through for 5 years or 50,000 miles.† See your dealer for details.

Now the LeBaron Series includes an American classic, the "woody" convertible. The newest way to enjoy high mileage, room, luxury and exceptional protection. All at a surprising price.

Coupe only

44 EST. Hwy. 28 EPA EST. MPG

Coupe base price

$8,514**

Chrysler

BUCKLE UP FOR SAFETY.

WE'VE RE-ENGINEERED THE AMERICAN LUXURY CAR.
THE NEW CHRYSLER CORPORATION QUALITY ENGINEERED TO BE THE BEST


WINE & FOOD
Cabernet Sauvignon
Cabernet Sauvignon is the unrivaled King of California wine grapes.
The distinction is well deserved for it was this magnificent varietal that first put California on the world map of wine. A native of Bordeaux, the wines produced of this grape have, for centuries, been among the world's finest and most enduring.
WINE
As with all heartier red wines, Cabernet Sauvignon must be properly aged to be at its best. When fully mature, a deep aromatic bouquet typically is followed by round, full flavors suggesting berries and herbs, warm earthy flavors, rich in varietal character. Velvety smooth in texture, it is a gentle giant of a wine.
Recognized for its excellence, our gold medal winning 1977 Proprietor's Reserve Cabernet Sauvignon is classic in its composition and sufficiently aged to be entirely enjoyable now.
FOOD
Traditionally, a wine as full bodied and flavorful as Cabernet Sauvignon is served with dishes of beef or venison, hearty stews, cheeses and rich pastas.
In our home, we enjoy Cabernet with a wide variety of foods. One I'm especially fond of is our own Eagle Ranch Roast, a wine country dish developed by my wife Vicki. Generously seasoned with olive oil, garlic and rosemary, this complex layering of flavors would overpower a wine of lesser body and character.
If you would enjoy sampling some of Vicki's recipes and learning more about wine, please write to the address below for your free brochure, "Wine and Food: Cabernet Sauvignon" a collection of family favorites uniquely compatible with this King of California Wines.
Discover Sebastiani Vineyards, a family winery in the Sonoma Valley.
Sam J. Sebastiani
THE RESULT OF IN THE VINEYARDS THIS WINE REPRESENTS TO EXCELLENCE
SPECIAL CARE AND AT THE WINERY, MY FIRM DEDICATION IN WINEMAKING.
PROPRIETOR'S RESERVE
VINEYARDS ESTABLISHED 1825
Sebastiani
1977
NORTH COAST COUNTIES
CABERNET SAUVIGNON
PRODUCED AND BOTTLED BY SEBASTIANI VINEYARDS
SONOMA, CALIFORNIA ALC. 13.1% BY VOL.
BONDED WINERY 876
Sebastiani Vineyards, Dept. G, P.O. Box AA, Sonoma, CA 95476
© 1982 SEBASTIANI VINEYARDS

Established 1845 **SCIENTIFIC AMERICAN** March 1983 Volume 248 Number 3

ARTICLES

DEPARTMENTS

SCIENTIFIC AMERICAN (ISSN 0036-8733), PUBLISHED MONTHLY BY SCIENTIFIC AMERICAN, INC., 415 MADISON AVENUE, NEW YORK, N.Y. 10017. COPYRIGHT © 1983 BY SCIENTIFIC AMERICAN, INC. ALL RIGHTS RESERVED. PRINTED IN THE U.S.A. NO PART OF THIS ISSUE MAY BE REPRODUCED BY ANY MECHANICAL, PHOTOGRAPHIC OR ELECTRONIC PROCESS, OR IN THE FORM OF A PHONOGRAPHIC RECORDING, NOR MAY IT BE STORED IN A RETRIEVAL SYSTEM, TRANSMITTED OR OTHERWISE COPIED FOR PUBLIC OR PRIVATE USE WITHOUT WRITTEN PERMISSION OF THE PUBLISHER. SECOND-CLASS POSTAGE PAID AT NEW YORK, N.Y., AND AT ADDITIONAL MAILING OFFICES. AUTHORIZED AS SECOND-CLASS MAIL BY THE POST OFFICE DEPARTMENT, OTTAWA, CANADA, AND FOR PAYMENT OF POSTAGE IN CASH. SUBSCRIPTION RATE: $21 PER YEAR IN THE U.S. AND ITS POSSESSIONS; $27 PER YEAR IN ALL OTHER COUNTRIES. POSTMASTER: SEND ADDRESS CHANGES TO SCIENTIFIC AMERICAN, 415 MADISON AVENUE, NEW YORK, N.Y. 10017.

TITURE, NOT DEMOLITION."

Chairman of the Board, AT&T

The task of divesting is vastly complicated. But we have the human talent. We have the engineering and scientific resources. And we have our pride.

While our companies will be radically changed, the basic aims of AT&T and the telephone companies will not change: to serve the communications needs of the nation, and to assure America its place as a leader in high technology worldwide.

Restructured by a Consent Decree with the Justice Department, the new AT&T will be able to develop and apply its technologies to the fullest.

In considering the Consent Decree proposal, we thought long and hard about whether the new structure would imperil our ability to serve. And of course we never would have accepted any terms that automatically would have degraded the quality of service to the nation. The Decree gives us the opportunity to build even better communications services and systems because it removes the arbitrary constraints which have limited us in the applications of our technologies.

We're eager to bring the benefits of all our innovations to the consumer in the competitive Information Age marketplace.

Both AT&T and the local phone companies have a bright future.

We're at the heart of the fastest growing, most promising industry in the country. The new AT&T—with its long-distance network and its new subsidiary, American Bell, plus Western Electric and Bell Labs—is in league with the future.

The telephone companies, which already provide a local communications system reaching virtually every home and business, are enhancing the quality and capability of their local lines so they can handle total information services: voice, video, data, even the new cellular mobile phone services. We are pledged to divest these companies in sound financial shape. We will keep that pledge.

This industry is where the opportunities and the excitement are.

If we have a truly competitive communications marketplace, with regulation only where it is needed, I believe AT&T and the divested telephone companies have a significant and constructive role to play in revitalizing the American economy and in maintaining and enhancing US technological leadership in communications.

The Bell System as we now know it will be no more. We will divest. But we are not demolishing the promise of tomorrow. That promise is alive and well. Bell System people are ready for new directions.

Bell System

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 75 cents each. Correspondence, orders and requests for the catalogue should be addressed to W. H. Freeman and Company, 4419 West 1980 South Salt Lake City, UT 84121. 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 $2 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 0036-8733/83. $2.00 + 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 they are accompanied by postage.

Advertising correspondence should be addressed to C. John Kirby, Advertising Director, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 10017.

Subscription correspondence should be addressed to Subscription Manager, SCIENTIFIC AMERICAN, P.O. Box 5969, New York, N.Y. 10017. The date of the last issue on your subscription is shown in the upper right-hand corner of each month's mailing label. For change of address notify us at least four weeks in advance. Please send your old address (if convenient, on a mailing label of a recent issue) as well as the new one.

Name ______________________

New Address ______________________

Street ______________________

City ______________________

State and ZIP ______________________

Old Address ______________________

Street ______________________

City ______________________

State and ZIP ______________________

THE COVER

The painting on the cover shows an experiment testing the memory of a marsh tit, a small British bird related to the North American chickadees (see "Memory in Food-hoarding Birds," by Sara J. Shettleworth, page 102). The bird is fitted with a hemispherical patch that covers one eye; then it is given sunflower seeds, which it hoards by poking each one into the moss that fills a large tray. Later the bird is allowed to search for the seeds. If the patch has not been moved, the bird remembers where the seeds are hidden. If the patch has been moved, however, so that it covers the eye on the opposite side of the head (the eye that viewed the storage sites), the result is quite different: the bird cannot remember. In the marsh tit each eye sends information to only one side of the brain. Thus it appears that visual memories in the marsh tit cannot be shunted from one side of the brain to the other. The experiment was done by David F. Sherry, John R. Krebs and Richard J. Cowie of the University of Oxford.

THE ILLUSTRATIONS

Cover painting by Enid Kotschnig

Page	Source
26–29	Edward Bell
31	Ilil Arbel
43	UN Photo
44–45	Andrew Tomko
46	Walken Graphics
48	Fred N. Spiess and Peter E. Lonsdale, Scripps Institution of Oceanography
49	Andrew Tomko
51	Hewlett-Packard Company (*top*), Ilil Arbel (*bottom*)
52	Zilog, Inc.
53	Internal Revenue Service, U.S. Department of the Treasury (*top*); Ilil Arbel (*bottom*)
54–57	Ilil Arbel
59	James P. Blair, courtesy of the National Museum of Pakistan
60	James P. Blair, courtesy of the National Museum of India (*top*); American Museum of Natural History (*bottom*)
61–65	Andrew Christie
66	McGuire Gibson, Oriental Institute, University of Chicago (*top*); Andrew Christie (*bottom*)
79	Ernst F. J. Van Bruggen, State University at Groningen (*top*); Annika C. Arnberg and Ernst F. J. Van Bruggen, State University at Groningen (*bottom*)
80–84	Bunji Tagawa
85	Annika C. Arnberg and Ernst F. J. Van Bruggen, State University at Groningen (*top*); Bunji Tagawa (*bottom*)
86–87	Bunji Tagawa
88	J. H. Hoeimakers, University of Amsterdam
91	Ann Savage, Royal Observatory Edinburgh, Coonabarabran, Australia
92–100	Walken Graphics
103–104	Tom Prentiss
105	Stephen B. Vander Wall, Utah State University
106–110	Tom Prentiss
113	R. F. Bonifield
114–122	Allen Beechel
123	R. F. Bonifield
125–135	Carol Donner
136	Ilil Arbel
139–145	Michael Goodman

Over 300 manufacturers sell their microcomputers with Microsoft software.

There's a reason.

Powerful simplicity. That's the concept behind every software program we write. Powerful programs that let you spend more time thinking about the problem...and less time thinking about the computer. That concept has earned us the confidence of over 300 microcomputer manufacturers. Starting with the very first microcomputer you could buy. And today, Microsoft® languages, operating systems and applications software are running on well over a million microcomputers. Worldwide.

Made for each other. Manufacturers literally build their computers around Microsoft software. That's no exaggeration. In fact, Microsoft frequently participates in the initial design and development of microcomputers. That's a major reason why Microsoft software runs so well on the majority of the world's 8- and 16-bit systems.

User-oriented. Microsoft is the only software supplier to offer a full range of compatible operating systems, languages, utilities and applications programs. If you're not a computer expert, here's what that means to you: Better programs. Programs that are not only more powerful, but easier to learn and use.

Better tools. We can't buy your trust. We have to earn it. With better products. Tools you can easily use to solve complex problems. We started with the first BASIC for the first microcomputer. Today, we offer a broad range of proven tools for microcomputers. Including SoftCard™ and IBM® RAMCard,™ products that materially increase the capabilities of Apple ® and IBM PC computers. Plus, business and management software such as Multiplan,™ the powerful electronic worksheet you can learn to use in just a few hours. Better tools. Because they're designed specifically for your computer.

Ask your Microsoft dealer. Most microcomputer manufacturers offer their systems with some Microsoft software, but you'll undoubtedly want more. Programs and languages that solve your specific problems. Programs for your industry, business or home. Programs that do more, yet ask less. Ask your computer software dealer for a demonstration. You'll see why more than 300 microcomputer manufacturers offer their systems with Microsoft software. The reason is powerful simplicity.

BETTER TOOLS FOR MICROCOMPUTERS

MICROSOFT™

MICROSOFT CORPORATION
10700 NORTHUP WAY
BELLEVUE, WASHINGTON 98004

Microsoft is a registered trademark, and Multiplan, SoftCard, RAMCard and the Microsoft logo are trademarks of Microsoft Corporation. Apple and the Apple logo are registered trademarks of Apple Computer, Inc. IBM and the IBM logo are registered trademarks of International Business Machines Corporation. The Radio Shack logo is a registered trademark of Radio Shack, a Division of Tandy Corporation. The Burroughs logo is a registered trademark of Burroughs Corporation. The Victor logo is a registered trademark of Victor Business Products. The Wang logo is a registered trademark of Wang Laboratories, Inc.

LETTERS

Sirs:

My article "The Galileo Affair" [SCIENTIFIC AMERICAN, August, 1982] has generated a continuing stream of responses, but only recently did Pierre Conway, O.P., remind me that I could have carried the account of the procedures against Galileo one day further. I ended the story on June 21, 1633, the last day on which the panel of cardinals received testimony, and neglected to mention the final sentencing and extended abjuration on June 22, 1633.

In a sense Napoleon is responsible for this lapse. In 1810 the Vatican Archives were hauled to France as part of the aggrandizement of the Bibliothèque Nationale. The records of the Galileo trial were a particularly important desideratum for the Napoleonic officials. When the Archives were returned in 1814, many documents were missing, including this important group. Eventually one of the three record books turned up: the slim volume containing the letters from Cardinal Bellarmine and Galileo's notarization of the testimony, all illustrated in my article. Thus the Vatican Archives today have no 17th-century record of the actual sentencing. A contemporary Italian translation exists in the Modena State Archives (although my request for a copy went unanswered); the Latin version was published in full by the Jesuit Giovanni B. Riccioli in his 1651 *Almagestum novum,* but in writing my article I had temporarily forgotten this source as well as the English translation in Giorgio de Santillana's *The Crime of Galileo.*

In the sentencing the Inquisition accused Galileo of a "vehement suspicion of heresy," and in the loyalty oath that Galileo was forced to read he said in part:

"After it had been notified to me that [this] doctrine was contrary to Holy Scripture, I wrote and printed a book in which I discuss this new doctrine already condemned and adduce arguments of great cogency in its favor without presenting any solution of these. Hence I have been pronounced by the Holy Office to be vehemently suspected of heresy, that is to say, of having held and believed that the Sun is the center of the world and immovable and that the Earth is not the center and moves.

"Therefore, desiring to remove from the minds of your Eminences, and of all Catholic Christians, this vehement suspicion justly conceived against me, with sincere heart and unfeigned faith I abjure, curse and detest the aforesaid errors and heresies and generally every other error and sect whatsoever contrary to the Holy Church, and I swear that in the future I will never again say or assert, verbally or in writing, anything that might furnish occasion for a similar suspicion regarding me; but, should I know any heretic or person suspected of heresy, I will denounce him to this Holy Office...."

From that moment onward the heliocentric doctrine could have been interpreted as heresy de facto by virtue of this legal ruling, in spite of the fact that no official decree had been issued; in this sense Galileo's trial could be considered a trial for heresy even though up until June 22, 1633, no public declaration to that effect had been made. Nevertheless, when Riccioli summed up the arguments 18 years after the event, he wrote that the doctrine that the earth moves "had been condemned as heresy, *or at least as erroneous*" (my italics), thereby expressing the ambiguity; meanwhile, in editions of the Index of prohibited books published in Spain, Copernicus' book was explicitly permitted....

My conclusions therefore stand that at least up until Galileo's sentencing the heliocentric doctrine had not been declared heretical, that he was tried more for disobeying orders than for heresy and that the Vatican today has a curiously restricted set of options in exonerating Galileo.

OWEN GINGERICH

Center for Astrophysics
Harvard College Observatory
Smithsonian Astrophysical Observatory
Cambridge, Mass.

Sirs:

As an addendum to the fine article by Carl Pomerance, "The Search For Prime Numbers" [SCIENTIFIC AMERICAN, December, 1982], the largest prime number known ($2^{44,497} - 1$) has been superseded by a larger one. The new prime ($2^{86,243} - 1$) was discovered in September, 1982, by David Slowinski, working with a Cray Research Cray-1 computer system. It was confirmed by us, working with a Control Data Corporation Cyber 205 computer. We found this new prime to be a "plus" Mersenne prime.

The confirmation is particularly important because it was made with a new transform multiplication technique that squares numbers in $P \log P$ time. Conventional multiplication approaches require P^2 time (where P is the P in $2^P - 1$). The speed obtained with this technique should enable prime searches in number ranges previously beyond the reach of existing computer technology.

STEPHEN K. MCGROGAN

CURT NOLL

Hayward, Calif.

SCIENCE/SCOPE

The first two satellites to be launched from NASA's space shuttle were placed in orbit for about one-third the cost of a conventional launch, thus saving their owners millions of dollars. This particular model of communications satellite, designed by Hughes Aircraft Company, is relatively inexpensive to launch because it sits upright and snugly folded in special cradles in the cargo bay. This feature saves money because launch costs are based on how much room a satellite takes up and how much it weighs. The cradle contains spring mechanisms that eject the satellite from the bay, after which rocket motors propel it into geostationary orbit. The drum-shaped spacecraft stands 9 feet tall when compacted. But when it reaches orbit, a telescoping solar panel deploys and the antenna unfolds, bringing the satellite's overall height to more than 21 feet.

A compact liquid-crystal light valve is designed to serve as a real-time light modulator for many optical data-processing and projection uses. The Hughes light valve uses liquid-crystal and thin-film technology to combine high input-light sensitivity and high image resolution with low voltage and power requirements. Uses include: graphics projection systems for large-screen displays, high-resolution vision for industrial robots, radar and sonar signal-processing, identification of moving objects, high-resolution spectral analysis of wide-band signals, and hybrid optical-digital processing systems.

Hughes spends about $8 million a year in support of colleges and universities, including help to alleviate the shortage of engineering faculty and modern laboratory equipment. For example, Hughes hires faculty consultants part-time during the school year and full-time in the summer. The program establishes a continuous technical interaction and supplements the pay of faculty members. Also, Hughes donates textbooks, software packages, and a wide variety of used equipment -- including electronic instrumentation and components.

The Smithsonian Institution is installing a new security system to monitor many facilities continuously. The Hughes system includes burglar alarms, fire-sensing devices, voice communications channels, and closed-circuit TV. It will let Smithsonian personnel control entrances and exits, and watch over areas open to visitors. A computer will collect and display information on TV monitors and printers at a central control station. Hughes previously installed a facilities management system at the Smithsonian's National Air and Space Museum. That system provides a wide range of exhibit monitor and control functions.

Hughes is seeking engineers to develop advanced systems and components for many different weather and communications satellites, plus the Galileo Jupiter Probe. Immediate openings exist in applications software development, data processing, digital subsystems test, microwave/RF circuit design, power supply design, digital communications, signal processing, spacecraft antenna design, system integration test and evaluation, and TELCO interconnection. Send your resume to Ray Bevacqua, Hughes Space & Communications Group, Dept. SE, Bldg. S/41, M.S. A300, P.O. Box 92191, Los Angeles, CA 90009. Equal opportunity employer.

Creating a new world with electronics

HUGHES

HUGHES AIRCRAFT COMPANY
EL SEGUNDO, CALIFORNIA 90245

50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

MARCH, 1933: "There is no question that the great American automobile industry has been hard hit by the general depression. The production of cars declined from 5,600,000 in 1929 to 3,500,000 in 1930, to 2,500,000 in 1931 and finally to fewer than 1,500,000 last year. Consequently, reckoning by the Department of Commerce's estimate of an average length of seven years for all cars, there should be an accumulated replacement demand for more cars in 1933 than were built in 1929. One noted, far-sighted automobile maker says he believes the point of saturation in per capita ownership of cars has not yet been reached. This, together with the present replacement demand and the fact that the foreign market is being more tightly closed against American cars, points to the necessity for concentration on the local market. By concentration, however, is not meant an efflorescence of ballyhoo and high-pressure salesmanship but the production of real and outstanding improvements."

"The next step in aircraft design—the conquest and exploitation of the so-called stratosphere—is close at hand. The two thrilling balloon flights of Professor Auguste Piccard and his aide into the stratosphere to make observations have done much to pave the way for power flights in that region. Once the problems of stratosphere flight are tackled, engineers and scientists will probably find a stratum where air travel will always be a pleasure because of the absence of 'bumps' and other atmospheric disturbances. This is the reason our airline operators will follow closely the exploration of the higher altitudes. The high-altitude ship may solve many present-day difficulties in connection with long distance non-stop flights on a profitable commercial basis. The expected higher speeds in the thin air, with sea-level horsepower and fuel load, will leave room for the pay load necessary to make transport operations worth while at lower rates and will make possible twice the present range of from 500 to 700 miles."

"Henry Norris Russell of Princeton University, whose monthly astronomical articles in this magazine are well known to our readers, has just been elected to the presidency of the American Association for the Advancement of Science, the largest of all organized bodies of general scientific men. The A.A.A.S. has almost 20,000 members and comprises practically the entire personnel of the many branches of American science—mathematics, astronomy, physics, chemistry, geology, zoology and so on. It is unofficially the great general 'holding company,' federation or overhead organization of the several sciences, each of which has its own organization, and the award of its presidency is regarded by all men of science as a signal honor."

SCIENTIFIC AMERICAN

MARCH, 1883: "Last year 37,951 persons died in New York City, the ratio being a little over 29 per 1,000 of population. The number of deaths from the principal contagious diseases was as follows: smallpox, 269; measles, 912; scarlet fever, 2,070; diphtheria, 1,521; croup, 730; erysipelas, 151; typhus fever, 66; typhoid fever, 363; malarial fever, 533."

"The relations existing in matter during combustion under atmospheric pressure extend to those creative fires whose ashes form new worlds. On this hypothesis we are led to a chemical condition or period when compounds can begin to form. The combination would necessarily take place in the outer and cooler portions of a star's atmosphere. From this behavior of first principles or elements are derived not only all the substances—solids, liquids and gases—of the earth but also of all other forms of matter throughout the universe. Such are the world fires in the nebular theory of creation. We are, however, led back beyond this point into a gaseous chaos when the whole universe, inconceivable ages ago, was equally filled with a homogeneous mass of tenuous matter at an extremely high degree of temperature. The millions of bodies composing the different solar systems originated only in consequence of rotary movement during which a number of masses acquired greater density than the remaining gaseous mass and then acted upon the latter as central points of attraction."

"Consul E. L. Baker of Buenos Ayres sets forth the success that has attended the introduction of *Eucalyptus globulus* in the Argentine Republic. He has several times in his annual reports referred to the successful introduction of *Eucalyptus* (the blue gum tree of Australia) into the Argentine, and has spoken of the rapidity with which some portions of the pampas heretofore destitute of timber are now being dotted with plantations of these magnificent trees; from the ease with which they can be grown, and the marvelous rapidity of their growth, he has suggested the feasibility of their cultivation in the milder parts of the United States. From what Consul Baker has observed during his stay in the Argentine Republic he is more and more convinced that the eucalyptus is a most desirable tree with which to timber our Southwestern plains and renew our rapidly decreasing forests."

"Professor D. E. Hughes, F.R.S., has formed a theory of magnetism entirely based upon experimental results, and these have led him to the following conclusions: 1. Each molecule of a piece of iron, steel or magnetic metal is a separate and independent magnet, having its two poles and distribution of magnetic polarity exactly the same as its total evident magnetism when noticed upon a steel bar magnet. 2. Each molecule, or its polarity, can be rotated in either direction upon its axis by torsion, stress or physical forces such as magnetism and electricity. 3. The inherent polarity or magnetism of each molecule is a constant quantity like gravity; it can be neither augmented nor destroyed. 4. When we have external neutrality, or no apparent magnetism, the molecules or their polarities arrange themselves so as to satisfy their mutual attraction by the shortest path and thus form a complete closed circuit of attraction. 5. When magnetism becomes evident, the molecules or their polarities have all rotated symmetrically in a given direction, producing a North Pole if rotated in this direction, as regards the piece of steel, or a South Pole if rotated in the opposite direction."

"The *Journal des Usines à Gaz,* on the subject of the use of ammoniacal liquor as a manure, states that it is so highly appreciated by Belgian agriculturists that the entire production of the gas works at Malines was bought up. Upon newly cleared ground the liquor was used just as it left the works, but for irrigation purposes it was diluted with three or four times its bulk of water. The effect produced on the soil by the use of the liquor is stated to be exactly the same as when stable dung is employed."

"The Coast and Geodetic Survey steamer *Blake* returned to New York on February 14 from a winter cruise for deep-sea exploration between the Bermudas and the Bahamas. On January 19, in latitude 19°41′ N, longitude 66°24′ W, about 105 miles northwest of St. Thomas, there was found the greatest depth ever measured in the Atlantic, or 4,561 fathoms. The place was about 80 miles southwest of the place where the *Challenger* made her deepest sounding of 3,862 fathoms."

4

Fourth in a series on how Delco Electronics and Bose technology contribute to your enjoyment of driving.

"...you have to hear it to believe it!"
Popular Mechanics

"A new and revolutionary sound system has been developed by the most unlikely partnership ever to be created in the audio industry...General Motors...has teamed up with Bose Corporation to create a car stereo system that is so far ahead of anything currently available in car audio that audio enthusiasts who can afford it may well be spending more time listening to music in their cars than they do at home."
Len Feldman, Audio Times

"The performance of the Delco-GM/Bose Music System was astounding...I can't imagine anyone...buying (one of these cars) without the Music System."
Gary Stock, High Fidelity

"If your car is this well equipped, you won't want to go home again."
Rich Warren, Chicago Magazine

The Delco-GM/Bose Music System is available as a factory installed option on Cadillac Seville and Eldorado, Buick Riviera, and Oldsmobile Toronado.


709·XNY
NEW YORK

The Mercedes-Benz 300 SD Turbodiesel is one of the most expensive, least extravagant corporate automobiles you can buy.

The 300 SD is $38,000* worth of automotive enlightenment–a corporate flagship meant to stress not pomp but efficiency. Meanwhile, its retained value over the past three years has been shown to *average* 90 percent.

Does your company have the technology of the 1990's in its executive offices, and the technology of the 1950's in its executive garage?

The fact is that inefficient operation, excessive size and savage depreciation need *not* define today's corporate flagship. The 300 SD Turbodiesel represents a different and refreshingly more sane equation.

10,000 miles @ $380?

The 300 SD's basic design efficiencies permit the efficiency of a diesel engine only *three liters* in size. It is no ponderous V-8 but an in-line five–and such a light drinker that a fuel cost of about $380 per annum is conceivable, based on 10,000 miles' driving, current average diesel fuel prices of $1.25 per gallon, and a 33 mpg EPA highway mileage figure. (City mpg [27] EPA.)**

The dollars and cents may not be crucial; the concept of relentless efficiency that such figures reflect, however, is as bracing as it is novel in the world of company cars.

So sanguine is Mercedes-Benz about the reliability of the 300 SD and its engine, incidentally, that it comes with a 36-month-or-36,000-mile warranty.†

This frugal diesel is meanwhile the most *powerful* such engine yet placed in a production automobile. Old images of "diesel lag" are obliterated in a turbinelike rush of energy and brisk acceleration, generated by a built-in turbocharger.

Five adults will find themselves extremely well cared for aboard the 300 SD. Its interior is almost 109 cubic feet worth of first-class repose, outfitted with only first-class amenities.

Yet so disciplined is its design that the car is less than 17 feet long and turns within just 39 feet. At 3,780 pounds, it is neither ponderous nor flimsy but athletically trim.

Its all-welded body achieves remarkable solidity and strength in part by using high strength, low alloy steels. Both the trunk lid and hood are fabricated in aluminum–part of the engineers' relentless quest to pare off weight wherever possible.

Not for boulevards only

If the 300 SD doesn't perform like the usual diesel, neither does it handle like the usual limousine in the clutch.

Mercedes-Benz, refreshingly, assumes that even company presidents must sometimes face switchback curves, slippery spots and potholes. The 300 SD is girded for such adversity: its fully independent suspension, zero-offset front suspension geometry and forged light-alloy wheels may be less showy than opera windows or carriage lamps.

There is no real need for a company driver to pilot the 300 SD. Its acute precision of response makes it deeply pleasurable to drive, across town or across country. A "driver's car," defined. For instance, note that its crisp four-speed automatic gearbox can also be shifted by hand.

120 safety features

Passengers are hardly ignored. Twin reading lamps are recessed in the rear of the cabin. There is even a separate ventilation console to serve the rear-seat occupants. The complement of standard features includes electronic cruise control, electric window lifts and front-seat adjustment, AM/FM stereo radio/cassette player with four speakers, and trimming in genuine hand-finished woods. No fewer than 120 safety features are also standard.

The 300 SD's history of value retention outstrips that of any luxury sedan extant. In fact, the *N.A.D.A. Official Used Car Guides* for 1982 calculate an average retained value for the 300 SD, over the past three years, of 90 percent. This helps place its $38,000 price in proper perspective.

A persuasive statement

There is one final argument for bringing a Mercedes-Benz 300 SD Turbodiesel Sedan into the firm. While efficiently serving the firm, it serves also as a fine advertisement–identifying your company as progressive-minded, and sensible, and shrewd. What limousine has ever made a statement quite like that?

*Approximate suggested advertised delivered price at port of entry. **EPA estimate for comparison purposes. The mileage you get may vary with trip length, speed and weather. †This is, of course, a limited warranty; you should consult your authorized Mercedes-Benz dealer for full details.

Engineered like no other car in the world

3 Professional Engineering Book Clubs

SAVE UP TO $81.00
Off Publishers' Prices When You Join with this Offer!

Macmillan has three nationally-known Professional Engineering Book Clubs designed for continuing education and career growth: one for Electrical and Electronics Engineers, one for Civil Engineers and one for Mechanical Engineers. If you're a practicing engineer—rely on Macmillan to help you maintain a competitive edge that allows you to move ahead in your job and excell in your chosen field of engineering.

Pub. Ed. Price $79.95

Only $2.95

when you join the

PROFESSIONAL MECHANICAL ENGINEERING BOOK CLUB

- 11th Edition, revised and expanded
- Over 1,700 pages
- 20 all-inclusive sections
- Thousands of charts, tables and schematics
- Covers important areas such as Units, Symbols and Constants, Theory and Design, Metallurgy and Non-Metals Technology, Welding and Surface Finishes, Non-Destructive Testing, Lubricants, Instrumentation, Plastics, Computers in Design, Thermodynamics, Fluid Mechanics, and much more.

3 **The Professional Mechanical Engineering Book Club** Now get the books you need to get ahead in mechanical engineering—definitive books that deal with subjects like materials and processes in manufacturing; measurements and control applications; reliability; mini and microcomputer controls in the industrial process; automation; pressure vessel design, and much more.

See below for club details, then use the card bound into this ad to join the Professional Mechanical Engineering Book Club and get your *Mechanical Engineer's Reference Book, 11th Edition* for only *$2.95* (Publisher's Price: $79.95).

FOUR GOOD REASONS TO JOIN

1. The Finest Books. Of the hundreds and hundreds of books submitted to us each year, only the very finest are selected and offered. Moreover, our books are always of equal quality to publishers' editions, *never* economy editions.

2. Big Savings. In addition to getting the **Engineer's Reference Book of your choice** for only $2.95 when you join, you keep saving substantially—up to 30% and occasionally even more. (For example, your total savings as a trial member, including this introductory offer, can easily be over 50%. That's like getting every other book free!)

3. Bonus Books. Also, you will immediately become eligible to participate in our Bonus Book Plan, with savings of up to 70% off the publishers' prices.

4. Convenient Service. At 3-4 week intervals (16 times per year) you will receive the Book Club News, describing the Main Selection and Alternate Selections, together with a dated reply card. If you want the Main Selection, do nothing and it will be sent to you automatically. If you prefer another selection or no book at all, simply indicate your choice on the card, and return it by the date specified. You will have at least 10 days to decide. If, because of late mail delivery of the News, you should receive a book you do not want, we guarantee return postage.

If reply card has been removed, please write to Macmillan Professional Engineering Book Club, Dept.—AK6, Riverside, N.J. 08075 to obtain membership information and an application.

THE AUTHORS

ELISABETH MANN BORGESE ("The Law of the Sea") is professor of political science at Dalhousie University in Nova Scotia. The daughter of the novelist Thomas Mann, she was born in Munich and received a diploma from the Conservatory of Music in Zurich in 1937. In 1938 she emigrated to the U.S. with her family, becoming a naturalized citizen in 1941. Before she joined the faculty at Dalhousie in 1978 she was for many years a senior fellow at the Center for the Study of Democratic Institutions in Santa Barbara. The law and politics of the oceans are her major interests. She is the author of many articles and books on the subject, including *The Ocean Regime* (1968) and *Seafarm: The Story of Aquaculture* (1980).

DAVID A. PATTERSON ("Microprogramming") is associate professor of electrical engineering and computer science at the University of California at Berkeley. Before earning his doctorate in 1976, he worked for the Hughes Aircraft Company designing and evaluating computers. He has been at Berkeley since 1977. He spent the fall of 1979 on leave at the Digital Equipment Corporation developing techniques for the design of microprograms. In one of his courses at Berkeley he and his students designed a microprocessor called RISC I (for Reduced-Instruction-Set Computer I), in which all programs in effect become microprograms.

WALTER A. FAIRSERVIS, JR. ("The Script of the Indus Valley Civilization"), is professor of anthropology at Vassar College. He received an associate of arts degree from the University of Chicago in 1953. He went on to obtain his bachelor's degree at Columbia University and his doctorate in anthropology from Harvard University. His professional career has been divided between teaching and leading anthropological expeditions to the Middle East and Far East. He has led expeditions to Afghanistan, Pakistan, Iran and India, some of them under the auspices of the American Museum of Natural History, of which he is a research associate. He held academic appointments at New York University and the University of Washington before moving to Vassar in 1968. Fairservis writes: "I have an implicit belief that in the origins of civilization we are dealing with something we do not completely understand. My whole career has been aimed at increasing our understanding of that."

LESLIE A. GRIVELL ("Mitochondrial DNA") was born and educated in England, receiving his B.S. in 1966 and his Ph.D. in biochemistry in 1969 from University College London. He writes: "My interest in mitochondrial biogenesis was aroused during my undergraduate studies and I was particularly fortunate to enter the field at the time of its upswing in the late 1960's and early 1970's." Since 1969 he has been at the Laboratory of Biochemistry of the University of Amsterdam, where he is currently professor of macromolecular biology. Outside the laboratory he describes himself as an "enthusiastic although somewhat inexpert musician. The instruments I play (recorders, flute and clarinet) reflect phases of interest that have ranged from the Renaissance through baroque and classical periods to the romantics."

DUANE A. DICUS, JOHN R. LETAW, DORIS C. TEPLITZ and VIGDOR L. TEPLITZ ("The Future of the Universe") are physicists who share an interest in the relations between high-energy physics and cosmology. Dicus is associate professor of physics at the University of Texas at Austin. His bachelor's and master's degrees are from the University of Washington and his doctorate in physics is from the University of California at Los Angeles. After serving as research associate at U.C.L.A., the Massachusetts Institute of Technology and the University of Rochester he moved to the University of Texas in 1973. Letaw is technical director of the Severn Communications Corporation in Severna Park, Md. He received his B.A. at Clark University in 1975 and his Ph.D. in physics from the University of Texas at Austin in 1981. Doris Teplitz is currently visitor at the University of Maryland at College Park. Her B.S. is from Wellesley College and her Ph.D. in physics is from Northeastern University. From 1974 to 1978 she was a member of the faculty of the Virginia Polytechnic Institute and State University. Vigdor Teplitz is deputy chief of the Strategic Affairs Division of the U.S. Arms Control and Disarmament Agency and adjunct professor of physics at the University of Maryland. His B.S. is from M.I.T.; his Ph.D. in physics is from the University of Maryland. Before taking up his current jobs he held appointments at M.I.T. and Virginia Polytechnic Institute and State University. In 1979 he served as adviser to the U.S. delegation at the second and third sessions of the negotiations on antisatellite weapons between the U.S. and the U.S.S.R.

SARA J. SHETTLEWORTH ("Memory in Food-hoarding Birds") is professor of psychology at the University of Toronto. She was graduated from Swarthmore College in 1965 with a B.A. and went on to get her M.A. in 1966 from the University of Pennsylvania and her Ph.D. in psychology in 1970 from the University of Toronto. She writes: "My Ph.D. thesis describes how reward and punishment have different effects on different activities such as grooming, digging and rearing in golden hamsters. More recent interests include the role of learning and memory mechanisms in foraging behavior. I have also done work on sea-turtle behavior with my husband Nicholas Mrosovsky."

IRVING R. EPSTEIN, KENNETH KUSTIN, PATRICK DE KEPPER and MIKLÓS ORBÁN ("Oscillating Chemical Reactions") are chemists with a common interest in the subject of their article. Epstein is professor of chemistry at Brandeis University. He holds three degrees from Harvard: an A.B. from Harvard College and an M.A. and a Ph.D. from Harvard University. He went to Brandeis in 1971. In 1977 and 1978 he was a National Science Foundation fellow at the Max Planck Institute for Biophysical Chemistry in Göttingen. Kustin is also professor of chemistry at Brandeis. He received his B.S. from Queens College in 1955. His Ph.D. in inorganic chemistry is from the University of Minnesota. After serving as U.S. Public Health Service fellow at the Max Planck Institute for Physical Chemistry in Berlin he moved to Brandeis in 1961. From 1974 to 1977 he served as chairman of the department of chemistry there. De Kepper is a manager of research in chemical engineering at the French National Center for Scientific Research (CNRS) Paul Pascal Research Center at Domaine University in Talence. A native of France, he was educated at the University of Bordeaux I, earning his Ph.D. in chemical engineering in 1978 and going on to the CNRS laboratory in the same year. Orbán was born and educated in Hungary. His Ph.D. is from Eötvös University in Budapest, where he has been a member of the faculty since 1962; he is currently associate professor of chemistry.

STEPHEN A. GELLER ("Autopsy") is professor of clinical pathology and vice-chairman of the department of pathology at the Mount Sinai School of Medicine. He obtained his undergraduate education at Brooklyn College and went on to earn his M.D. in 1964 from the College of Medicine of Howard University. He was an intern at Lenox Hill Hospital in New York City and resident in pathology at Mount Sinai. He served for two years as pathologist at the Naval Hospital in Beaufort, S.C., before proceeding to Mount Sinai to take up his current work. His specialty is the diseases of the gastrointestinal tract.

METAMAGICAL THEMAS

Tripping the light recursive in Lisp, the language of artificial intelligence

by Douglas R. Hofstadter

Since I ended last month's column with a timely newsbreak about the homely Glazunkian porpuquine, I felt it only fitting to start off this month's with more about that little-known but remarkable beast. As you may remember, the quills on any porpuquine (except the tiniest ones) are smaller porpuquines. The tiniest porpuquines have no quills but do have a nose, and a very important nose at that, since the Glazunkians base their entire monetary system on that little object. Consider the value of three-inch porpuquines in Outer Glazunkia. Each porpuquine always has nine quills (contrasting with its cousins in Inner Glazunkia, which always have seven); thus each has nine two-inch porpuquines sticking out of its body. Each of those in turn sports nine one-inch porpuquines, out of each of which sprout nine zero-inch porpuquines, each of which has one nose. All told this comes to $9 \times 9 \times 9 \times 1$ noses, which means that a three-inch porpuquine in Outer Glazunkia has a buying power of 729 noses. If, in contrast, we had been in Inner Glazunkia and had started with a four-incher, that porpuquine would have a buying power of $7 \times 7 \times 7 \times 7 \times 1$, or 2,401, noses.

Let us see if we can come up with a general recipe for calculating the buying power (measured in noses) of any porpuquine. It seems to me that it would go something like this:

The buying power of a porpuquine with a given quill count and size is:
- if its size equals 0, then 1;
- otherwise figure out the buying power of a porpuquine with the same quill count but of the next-smaller size and multiply that by the quill count.

We can shorten this recipe by adopting some symbolic notation. First, let "q" stand for the quill count and "s" for the size. Then let "cond" stand for "if" and "t" for "otherwise." Finally, use a kind of condensed algebraic notation in which the English names of operations are placed to the left of their operands, inside parentheses. We get something like this:

```
(buying-power q s) is:
    cond (eq s 0) 1;
         t (times
               q
               (buying-power
                   q
                   (next-smaller s)))
```

This is an exact translation of the earlier English recipe into a slightly more symbolic form. We can make it a little more compact and symbolic by adopting a couple of new conventions. Let each of the two cases (the case where s equals 0 and the "otherwise" case) be enclosed in parentheses; in general, use parentheses to enclose each logical unit completely. Finally, indicate by the words "def" and "lambda" that this is a definition of a general notion called "buying power" with two variables (quill count q and size s). Now we get:

```
(def buying-power
 (lambda (q s)
  (cond ((eq s 0) 1)
        (t
         (times q
           (buying-power q
              (next-smaller s)))))))
```

We mentioned above that the buying power of a nine-quill, three-inch porpuquine is 729 noses. This could be expressed by saying that (buying-power 9 3) equals 729. Similarly, (buying-power 7 4) equals 2,401.

Now let us get back to Lisp. I had posed a puzzle toward the end of last month's column in which the object was to write a Lisp function that subsumed a family of functions called "square", "cube", "4th-power", "5th-power" and so on. I asked you to come up with one *general* function called "power", having two variables, such that "(power 9 3)" gives 729, "(power 7 4)" gives 2,401 and so on. I had presented a "tower of power," that is, an infinitely tall tower of separate Lisp definitions, one for each power, connecting it to the preceding power. Thus a typical floor in this tower would be:

```
(def 102nd-power
 (lambda (q)
  (times q (101st-power q))))
```

Of course, 101st-power would refer to 100th-power in its definition, and so on, thereby creating a rather long regress back to the "embryonic," or simplest, case. Incidentally, that very simplest case, rather than "square" or even "1st-power", is this:

```
(def 0th-power (lambda (q) 1))
```

I told you that you had all the information necessary to assemble the proper definition. All you needed to observe is, of course, that each floor of the tower rests on the "next-smaller" floor (except for the bottom floor, which is a "stand-alone" floor). By "next-smaller" I mean the following:

```
(def next-smaller
 (lambda (s) (difference s 1)))
```

Thus "(next-smaller 102)" yields 101. Actually Lisp has a standard name for this, namely "sub1", and a name for its inverse as well, namely "add1". If we put all our observations together, we come up with the following universal definition:

```
(def power
 (lambda (q s)
  (cond ((eq s 0) 1)
        (t
         (times q
           (power q
              (next-smaller s)))))))
```

This is the answer to the puzzle I posed. H'm, that's funny. I have the strangest sense of déjà vu. I wonder why?

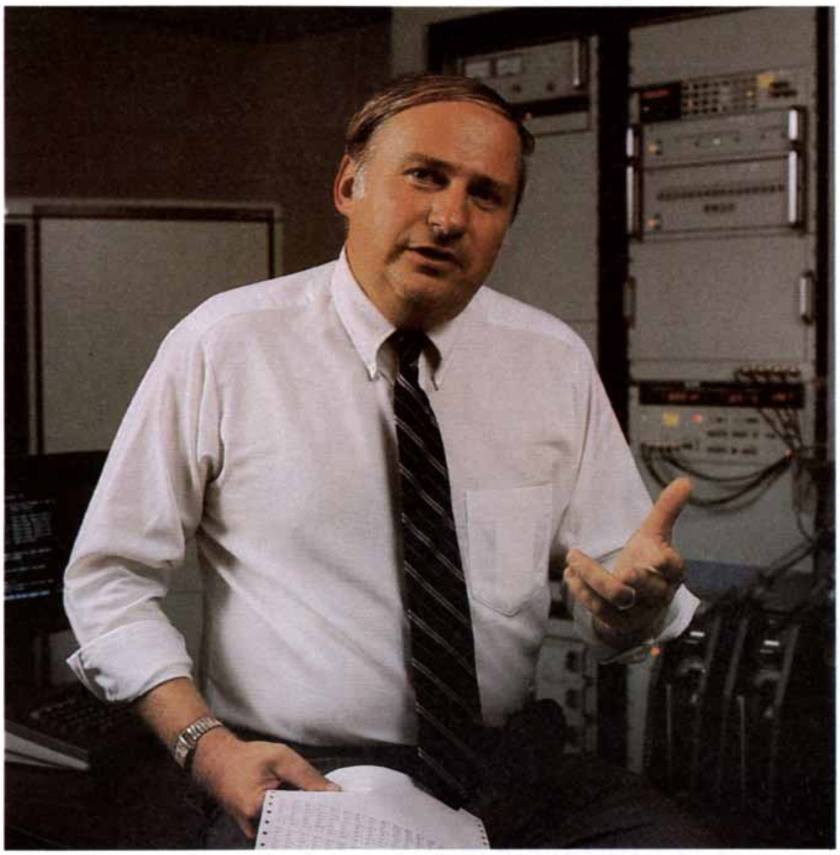
The definition presented here is called a *recursive* definition, for the reason that inside the definiens the definiendum is used. This is a fancy way of saying that I appear to be defining something in terms of itself, which ought to be considered gauche if not downright circular in anyone's book. To see whether the Lisp genie looks askance on such trickery let us ask it to figure out (power 9 3):

```
—> (power 9 3)
729
—>
```

Well, fancy that! No complaints? No choking? How can the Lisp genie swallow such nonsense?

The best explanation I can give is to

What if you chose as a technical



“At M/A-COM, HP computers increase test speed 12-fold.”

M/A-COM, Inc., is a major manufacturer of electronic products for commercial telecommunications and defense markets. Dr. Alan Carlson, Manager of the Automatic Test Equipment Group in Burlington, Mass., says: “We build automatic test systems and conduct final test on hundreds of M/A-COM products. To automate, we needed precision instruments plus computer power.

“HP was the only vendor to offer us a complete solution. With 20 HP 1000 computers driving more than 200 HP instruments, we've completely automated several product tests. And we can depend on HP's support teams to keep our entire testing network up and running.

“Now we can test sophisticated products 12 times faster than before. With these kinds of savings, each HP 1000 has paid for itself in less than one year.”

What should you expect from a computer partner?

Quality—from researching customer needs to product development, manufacturing, marketing, after-sales service, and support!

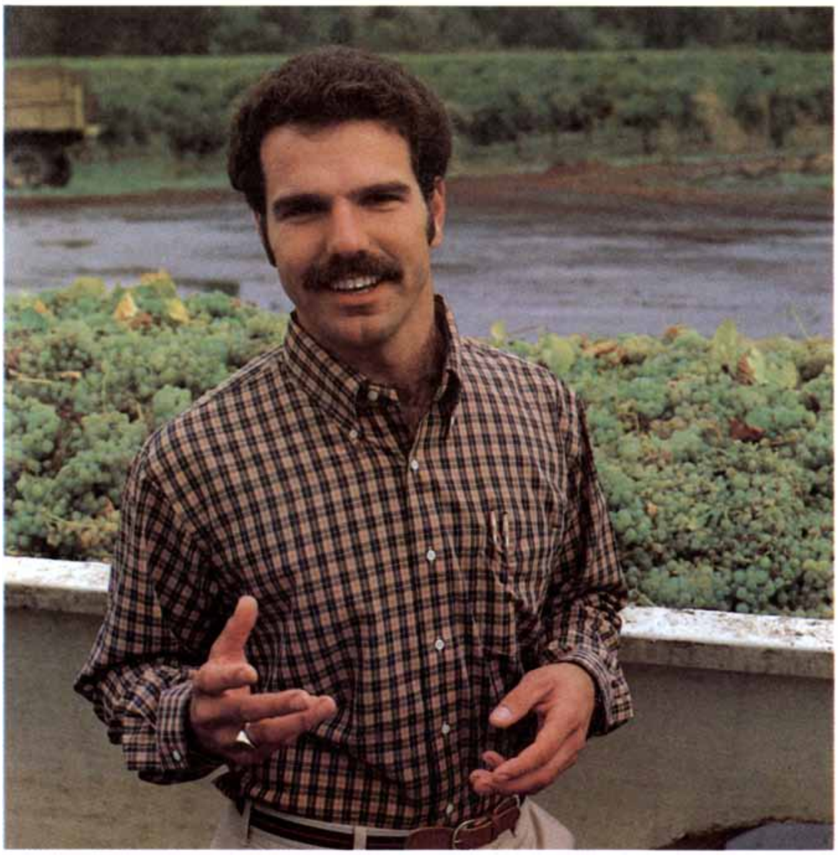
Quality, by HP's definition, encompasses more than the product.

This high quality recently was recognized by the Union of Japanese Scientists & Engineers, who awarded a coveted Deming Prize to Yokogawa-Hewlett-Packard, HP's joint venture company in Japan.

YHP offers the entire spectrum of HP products and in 1981 accounted for Japanese sales of more than $200 million. Nearly 60 percent of the products YHP sells are imported from HP manufacturing operations

***The Deming Prize:** for outstanding merit in the control of quality.*

Hewlett-Packard computer partner?



"At Charles Krug Winery, an HP computer tests temperature 150 times faster than manual methods, improves quality control, and doubles production."

Charles Krug and CK Mondavi premium wines are produced at the Charles Krug Winery in St. Helena, California. The winery uses an HP 1000 computer system to monitor and control the temperature of its fermentation tanks.

Treasurer Peter Mondavi, Jr. explains: "The HP 1000 tests the temperature of each tank 150 times faster than manual methods. And because it automatically adjusts fluctuating temperatures we know each bottle of wine will meet our quality standards.

"Improved temperature control has enabled us to double our production while maximizing quality control. As a result, we've saved $50,000 in production costs this year alone."

in the US and abroad. Conversely, YHP manufactures and exports various products to HP customers throughout the world.

The total quality control approach for which YHP was awarded the Deming Prize is a point of pride throughout HP. Customers in every part of the world should expect *quality* as a fundamental element in a computer partnership with HP.

You should consider a working partnership with HP. Now.

For a free copy of our brochure explaining HP's practical, proven approach to meeting your long-term computer and information needs, write to A.P. Oliverio, Senior Vice President, Marketing, Dept. 207, Hewlett-Packard Company, P.O. Box 10301, Palo Alto, CA 94303.

When performance must be measured by results

HEWLETT PACKARD

0002340

point out that no circularity is actually involved. Although it is true that the definition of "power" uses the word "power" inside itself, the two occurrences are referring to different circumstances. In a nutshell, (power q s) is being defined in terms of a simpler case, namely (power q (next-smaller s)). Thus I am defining the 44th power in terms of the 43rd power, and that in terms of the next-smaller power, and so on down the line until we come to the "bottom line," as I call it: the 0th power, which needs no recursion at all. It suffices to tell the genie that its value is 1, which we did. So when you look carefully, you see that this recursive definition is no more circular than the "tower of power" was—and you cannot get any straighter than an infinite straight line! In fact, this one compact definition really is just a way of getting the entire tower of power into one finite expression. Far from being circular, it is just a handy summary of infinitely many different definitions, all belonging to one family.

In case you still have a trace of skepticism about this sleight of hand, perhaps I should let you watch what the Lisp genie will do if you ask for a "trace" of the function and then ask it once again to evaluate (power 9 3). For this you will need to refer to the illustration below.

On the lines marked "ENTERING" the Lisp genie prints the values of the two arguments, and on the lines marked "EXITING" it prints the value it has computed and is returning. For each ENTERING line there is of course an EXITING line, and the two are aligned vertically, that is, they are indented by the same amount.

You can see that in order to figure out what (power 9 3) is, the genie must first calculate (power 9 2). This, however, is not a given; instead it requires knowing the value of (power 9 1), and this in turn requires (power 9 0). Ah! We *were* given this one: it is just 1. And now we can bounce back "up," remembering that in order to get one answer from the "deeper" answer we must multiply by 9. Hence we get 9, then 81, then 729, and we are done.

I say "we," but of course it is not we but the Lisp genie who must keep track of these things. The Lisp genie has to be able to suspend one computation to work on another one whose answer was asked for by the first one. And the second computation too may ask for the answer to a third one, thus putting itself on hold, as may the third, and so on recursively. Eventually, however, there will come a case where the buck stops—that is, where a process runs to completion and returns a value—and this will enable other stacked-up processes to finally return values, like stacked-up airplanes that have circled for hours finally getting to land, with each landing opening the way for another landing.

Ordinarily the Lisp genie will not print out a trace of what it is thinking unless you ask for it. Whether you ask to see it or not, however, this kind of thing is going on behind the scenes whenever a function call is evaluated. One of the enjoyable things about Lisp is that it can deal with such recursive definitions without getting flustered.

I am not so naive as to expect that you have now totally got the hang of recursion and could go out and write huge recursive programs with ease. Indeed, recursion can be a remarkably subtle means of defining functions, and sometimes even an expert will have trouble figuring out the meaning of a complicated recursive definition. I therefore thought I would give you some practice in working with recursion.

Let me give a simple example based on this riddle: How do you make a pile of 13 stones? A silly answer would be: Put one stone on top of a pile of 12 stones. Suppose we want to make a Lisp function that will give us not a pile of 13 stones but a list consisting of 13 copies of the atom "stone" or, in general, n copies of that atom. We can base our answer on the riddle's silly-seeming yet correct recursive answer. The general notion is to build the answer for n out of the answer for n's predecessor. Build how? Using the list-building function cons, that's how. What is the embryonic case? That is, for which value of n does this riddle present absolutely no problem at all? That's easy: When n equals 0, our list should be empty, which means the answer is nil. We can now put our observations together as follows:

```
(def bunch-of-stones
 (lambda (n)
 (cond ((eq n 0) nil)
       (t
        (cons 'stone
           (bunch-of-stones
              (next-smaller n)))))))
```

Now let us watch the genie put together a very small bunch of stones (with "trace" on, just for fun). For this you will need to refer to the illustration on the opposite page.

This illustrates what is called "consing up a list." Now let us try another one. This one is an old chestnut of Lisp and indeed of recursion in general. Look at the definition and see whether you can figure out what it is supposed to do; then read on to see if you were right.

```
(def wow
 (lambda (n)
 (cond ((eq n 0) 1)
       (t
        (times n (wow (sub1 n)))))))
```

Remember, "sub1" means the same as "next-smaller". For a lark, why don't you calculate the value of (wow 100)? (If you ate your mental Wheaties this morning, try it in your head.)

It happens that Lisp genies often mumble out loud while they are executing wishes, and I just happen to have overheard this one as it was executing the wish "(wow 100)". Its soliloquy ran something like this:

"H'm...(wow 100), eh? Well, 100 surely isn't equal to 0, so I guess the answer has to be 100 times what it *would* have been had the problem been (wow 99). All right—now all I need to do is figure out what (wow 99) is. Oh, this is going to be a piece of cake! Let's see, is 99 equal to 0? No, seems not to be, so I guess the answer to *this* problem must be 99 times what the answer *would* have been had the problem been (wow 98). Let's see..."

At this point the author, having some pressing business at the bank, had to leave the happy genie and did not return until some milliseconds afterward. When he did so, the genie was just finishing up, saying:

"...and now I just need to multiply *that* by 100, and I've got my final answer. Easy as pie! I believe it comes out to be 93326215443944152681699238856266700490715968264381621468592963895217599999322991560894146397615651828625369792082722375825118521091686400000000000000000000000000—if I'm not mistaken."

Is that the answer you got, dear reader? No? Oh, I see where you went wrong.

```
—> (power 9 3)
    ENTERING power (q = 9, s = 3)
        ENTERING power (q = 9, s = 2)
            ENTERING power (q = 9, s = 1)
                ENTERING power (q = 9, s = 0)
                EXITING   power (value: 1)
            EXITING   power (value: 9)
        EXITING   power (value: 81)
    EXITING   power (value: 729)
729
—>
```

The Lisp genie evaluates (power 9 3)

It was in your multiplication by 52. Go back and try it again from that point on, and be a little more careful in adding up those long columns. I'm quite sure you'll get it right this time.

This "wow" function is ordinarily called "factorial"; n factorial is usually defined as the product of all the numbers from 1 through n. A recursive definition, however, looks at things a bit differently: speaking recursively, n factorial is simply the product of n and the preceding factorial. It reduces the given problem to a simpler one of the same type. That simpler one will in turn be reduced, and so on down the line, until you come to the simplest problem of that type, which I call the embryonic case or the bottom line. People speak, in fact, of a recursion "bottoming out."

A *New Yorker* cartoon from a few years back illustrates the concept perfectly. It shows a man of 50 or so holding a photograph of himself taken roughly 10 years earlier. In that photograph he is holding a photograph of himself 10 years earlier than *that*. And on it goes, until eventually it bottoms out—quite literally—in a photograph of a bouncy baby boy in his birthday suit (bottom in the air). This idea of recursive photographs catching you as you grow up is quite appealing—I wish my parents had thought of it. Compare it with the more famous infinite regress on the Morton Salt package, where the Morton Salt girl holds a box of Morton Salt with her picture on it. Since the girl in the picture is no younger, however, there is no bottom line and the regress is—at least theoretically—endless.

The recursive approach works when you have a family of related problems at least one of which is so simple that it can be answered immediately. This I call the embryonic case. (In the factorial example that is the "(eq n 0)" case, whose answer is 1.) Each problem (for instance, "What is 100 factorial?") can be viewed as a particular case of one general problem ("How do you calculate factorials?"). Recursion takes advantage of the fact that the answers to various cases are related in some logical way to one another. (For example, I could very easily tell you the value of 100 factorial if only someone would hand me the value of 99 factorial; all I need to do is multiply by 100.) You could say that the recursioneer's motto is "Gee, I could solve *this* case if only someone would magically hand me the answer to the case that is one step closer to the embryonic case." Of course, this motto presumes that certain cases are in some sense "nearer" to the embryonic case than others. In fact, it presumes that there is a natural pathway leading from any case through simpler cases all the way down to the embryonic case, a pathway whose steps are clearly marked.

```
—> (bunch-of-stones 2)
    ENTERING bunch-of-stones (n = 2)
        ENTERING bunch-of-stones (n = 1)
            ENTERING bunch-of-stones (n = 0)
            EXITING    bunch-of-stones (value: nil)
        EXITING    bunch-of-stones (value: (stone))
    EXITING    bunch-of-stones (value: (stone stone))
(stone stone)
—>
```

The genie puts together a very small bunch of stones

As it turns out, this is a most reasonable assumption to make in all kinds of circumstances. To spell out the exact nature of this recursion-guiding pathway you have to answer two Big Questions:

1. What is the embryonic case?
2. What is the relation between a typical case and the next-simpler case?

Now, actually both of these Big Questions break up into two subquestions (as befits any self-respecting recursive question), one concerning how you recognize where you are or how to move, the other concerning what the answer is at any given stage. Thus, spelled out more explicitly, our Big Questions are:

1*a*. How can you know when you have reached the embryonic case?
1*b*. What is the embryonic answer?

2*a*. From a typical case, how do you take exactly one step toward the embryonic case?
2*b*. How do you build this case's answer out of the "magically given" answer to the simpler case?

Question 2*a* concerns the nature of the *descent* toward the embryonic case, or bottom line. Question 2*b* concerns the inverse aspect, namely the *ascent* that carries you back up from the bottom to the top level.

In the case of the factorial the answers to the Big Questions are:

1*a*. The embryonic case occurs when the argument is 0.
1*b*. The embryonic answer is 1.

2*a*. Subtract 1 from the present argument.
2*b*. Multiply the "magic" answer by the present argument.

Notice how the answers to these four questions are all incorporated in the recursive definition of "wow".

Recursion relies on the assumption that sooner or later you will bottom out. One way to be *sure* you will bottom out is to have all the "descending," or simplifying, steps move in the same direction at the same rate, so that your pathway is quite obviously linear. For instance, it is obvious that by subtracting 1 repeatedly, you will eventually reach 0, provided you started with a positive integer. It is also obvious that by performing the list-shortening operation of "cdr" you will eventually reach nil, provided you started with a finite list. For this reason recursions using "sub1" or "cdr" to define their pathway of descent toward the bottom are commonplace. I shall present a cdr-based recursion below, but first I want to show a funny numerical recursion in which the pathway toward the embryonic case is anything but linear and smooth.

Consider the famous "3n + 1" problem, in which you start with any positive integer, and if it is even, you halve it; otherwise you multiply it by 3 and add 1. Let us call the result of this operation on n "(hotpo n)" (standing for "half or triple plus one"). Here is a Lisp definition of hotpo:

```
(def hotpo
 (lambda (n)
  (cond ((even n) (half n))
        (t (add1 (times 3 n))))))
```

This definition presumes that two other functions either have been or will be defined elsewhere for the Lisp genie, namely "even" and "half" ("add1" and "times" being, as I mentioned earlier, intrinsic parts of Lisp). Here are the missing definitions:

```
(def even
 (lambda (n) (eq (remainder n 2) 0)))

(def half (lambda (n) (quotient n 2)))
```

What do you think happens if you begin with some integer and perform hotpo over and over again? Take 7, for instance, as your starting point. Before you do the arithmetic guess at what kind of behavior might result.

As it turns out, the pathway is often surprisingly chaotic and bumpy. For instance, if we begin with 7, the process leads us to 22, then 11, then 34, 17, 52, 26, 13, 40, 20, 10, 5, 16, 8, 4, 2, 1, 4, 2, 1, 4, 2, 1,... Note that we wind up in a short loop. Suppose, therefore, we agree

```
—> (pathway-to-1 3)
ENTERING pathway-to-1 (tato = 3)
    ENTERING pathway-to-1 (tato = 10)
        ENTERING pathway-to-1 (tato = 5)
            ENTERING pathway-to-1 (tato = 16)
                ENTERING pathway-to-1 (tato = 8)
                    ENTERING pathway-to-1 (tato = 4)
                        ENTERING pathway-to-1 (tato = 2)
                            ENTERING pathway-to-1 (tato = 1)
                            EXITING pathway-to-1 (value: (1))
                        EXITING pathway-to-1 (value: (2 1))
                    EXITING pathway-to-1 (value: (4 2 1))
                EXITING pathway-to-1 (value: (8 4 2 1))
            EXITING pathway-to-1 (value: (16 8 4 2 1))
        EXITING pathway-to-1 (value: (5 16 8 4 2 1))
    EXITING pathway-to-1 (value: (10 5 16 8 4 2 1))
EXITING pathway-to-1 (value: (3 10 5 16 8 4 2 1))
(3 10 5 16 8 4 2 1)
—>
```

How the genie "thinks" when the trace feature is on

that if we ever reach 1, we have hit bottom and can stop. You might well ask, "Who says we *shall* hit 1? Is there a guarantee?" Indeed, if you haven't tried it out a bit, there is no obvious reason to suspect that you will hit 1. (In the case of 7 did you suspect it would happen?) Numerical experimentation, however, reveals a remarkable reliability to the process; it seems that no matter where you start, you always do hit 1 after a while. (Try starting with 27 if you want a real roller-coaster ride!)

Can you write a recursive function to reveal the pathway followed from an arbitrary starting point "down" to 1? Note that I say "down" advisedly, since many of the steps are in fact *up*. Thus the pathway starting at 3 would be the list "(3 10 5 16 8 4 2 1)". In order to solve this puzzle you need to go back and answer for yourself the two Big Questions of Recursion, as they apply here. Note:

```
(cond ((not (want help))
       (not (read further)))
      (t (read further)))
```

First, about the embryonic case. This is easy. It has already been defined as the arrival at 1, and the embryonic, or simplest possible, answer is the list "(1)", a tiny but valid pathway beginning and ending at 1.

Second, about more typical cases. What operation will carry us from typical 7 one step closer to embryonic 1? Certainly not the "sub1" operation. No, by definition it is the function hotpo itself that brings you ever "nearer" to 1—even when it carries you *up*. This teasing quality is of course the entire point of the example. What about 2*b*, how to recursively build a list documenting our wildly oscillating pathway? Well, the pathway belonging to 7 is got by tacking (that is, consing) 7 onto the shorter pathway belonging to (hotpo 7), or 22. After all, 22 is one step closer to being embryonic than 7 is!

These answers enable us to write down the desired function definition, using "tato" as our dummy variable—"tato" being a well-known acronym for "tato (and tato only)", which recursively expands to "tato (and tato only) (and tato (and tato only) only)", and so forth.

```
(def pathway-to-1
 (lambda (tato)
  (cond ((eq tato 1) '(1))
        (t
         (cons tato
               (pathway-to-1
                  (hotpo tato)))))))
```

Look at the way the Lisp genie "thinks" (as revealed when the trace feature is on). Here you should refer to the illustration above.

Notice the total regularity (the *V* shape) of the left margin of the trace diagram, in spite of the chaos of the numbers involved. Not all recursions are so geometrically pretty when they are traced. The reason is that some problems require that *more than one subproblem* be solved. As a practical, real-life example of such a problem, consider how you might go about counting up all the unicorns in Europe. This is certainly a nontrivial undertaking, yet there is an elegant recursive answer: Count up all the unicorns in Portugal (the "car" of Europe, metaphorically speaking), then count up all the unicorns in the other 30-odd countries of Europe (the "cdr" of Europe, to continue the metaphor) and finally add the two results together.

Notice how this spawns two smaller unicorn-counting subproblems, which in turn will spawn two subproblems each, and so on. For instance, how can one count all the unicorns in Portugal? It is easy: add the number of unicorns in the Estremadura region (Portugal's "car") to the number of unicorns in the rest of Portugal (Portugal's "cdr"). And how do you count up the unicorns in Estremadura (not to mention those in the remaining regions of Portugal)? By further breakup, of course. What, then, is the bottom line? Well, regions can be broken up into districts, districts into square kilometers, square kilometers into hectares, hectares into square meters—and we can handle each square meter without further breakup.

Although this may sound rather arduous, there really is no way to conduct a thorough census other than to traverse every single part on every level of the full structure you have, no matter how giant it may be. There is a perfect Lisp counterpart to this unicorn census: it is the problem of determining how many atoms there are inside an arbitrary list. How can we write a Lisp function called "atomcount" that will give us the answer 15 when it is shown the following strange-looking list (which we shall call "brahma")?

```
(((ac ab cb) ac (ba bc ac))
            ab
 ((cb ca ba) cb (ac ab cb)))
```

One method, expressed recursively, is exactly parallel to that for ascertaining the unicorn population of Europe. See if you can write it down.

The idea is this. We want to construct the answer—namely 15—out of the answers to simpler atom-counting problems. Well, it is obvious that one atom-counting problem simpler than (atomcount brahma) is (atomcount (car brahma)). Another is (atomcount (cdr brahma)). The answers to these two problems are respectively 7 and 8. Now, clearly 15 is made out of 7 and 8 by addition—which makes sense, after all, since the total number of atoms must be the number in the car plus the number in the cdr. There is nowhere else for any atoms to hide. This analysis gives us the following recursive definition, with "s" as the dummy variable:

```
(def atomcount
 (lambda (s)
  (plus (atomcount (car s))
        (atomcount (cdr s)))))
```

It looks simple, but it has a couple of flaws. First, we have written the *recursive* part of the definition, but we have utterly forgotten the other equally vital half—the bottom line. It reminds me of a Maryland judge I once read about in the paper, who ruled that "a horse is a four-legged animal that is produced by two other horses." This is a lovely definition, but where does it bottom out? It is the same for atomcount. What is the simplest case, the embryonic case, of atomcount? It is when we are asked to count the atoms in a single atom. The answer in such a case is of course 1. But how can we know when we are looking at an atom? Fortunately Lisp has a built-in function called "atom" that returns t (meaning "true") whenever we are looking at an atom, and returns nil otherwise. Thus "(atom 'plop)" returns t and "(atom '(a b c))" returns nil. Using that, we can patch up our definition:

```
(def atomcount
 (lambda (s)
  (cond ((atom s) 1)
        (t
         (plus (atomcount (car s))
               (atomcount (cdr s)))))))
```

It is still, however, not quite right. If we ask the genie for the atomcount of "(a b c)", instead of getting 3 for an answer we shall get 4. Shocking! How does this happen? Well, we can pin the problem down by trying an even simpler example: if we ask for (atomcount '(a)), we find we get 2 instead of 1. Now the error should be clearer: 2 = 1 + 1, with 1 each coming from the car and the cdr of "(a)". The car is the atom "a", which indeed should be counted as 1, but the cdr is nil, which should not. Then why does nil give an atomcount of 1? Because nil is not only an empty list but also an atom! To suppress this bad effect we simply insert another cond clause at the very top:

```
(def atomcount
 (lambda (s)
  (cond ((null s) 0)
        ((atom s) 1)
        (t
         (plus (atomcount (car s))
               (atomcount (cdr s)))))))
```

I wrote "(null s)", which is just another way of saying "(eq s nil)". In general if you want to determine whether the value of some expression is nil or not, you can use the inbuilt function "null", which returns t if yes and nil if no. Hence, for example, "(null (null nil))" evaluates to nil, since the inner function call evaluates to t, and t is not nil! What about "(null '(null nil))"?

Now, what happens when we run our function on brahma, its original target? The result, with trace on, is shown below.

Notice the more complicated topography of this recursion, with its ins and outs along the left margin. The preceding *V*-shaped recursion looked like a

```
—>(atomcount brahma)
  ENTERING atomcount
     (s = (((ac ab cb) ac (ba bc ac)) ab ((cb ca ba) cb (ac ab cb))))
    ENTERING atomcount (s = ((ac ab cb) ac (ba bc ac)))
      ENTERING atomcount (s = (ac ab cb))
        ENTERING atomcount (s = ac)
        EXITING    atomcount (value: 1)
        ENTERING atomcount (s = (ab cb))
          ENTERING atomcount (s = ab)
          EXITING    atomcount (value: 1)
          ENTERING atomcount (s = (cb))
            ENTERING atomcount (s = cb)
            EXITING    atomcount (value: 1)
            ENTERING atomcount (s = nil)
            EXITING    atomcount (value: 0)
          EXITING atomcount (value: 1)
        EXITING atomcount (value: 2)
      EXITING    atomcount (value: 3)
      ENTERING atomcount (s = (ac (ba bc ac)))
        ENTERING atomcount (s = ac)
        EXITING    atomcount (value: 1)
        ENTERING atomcount (s = ((ba bc ac)))
          ENTERING atomcount (s = (ba bc ac))
            ENTERING atomcount (s = ba)
            EXITING    atomcount (value: 1)
            ENTERING atomcount (s = (bc ac))
              ENTERING atomcount (s = bc)
              EXITING    atomcount (value: 1)
              ENTERING atomcount (s = (ac))
                ENTERING atomcount (s = ac)
                EXITING    atomcount (value: 1)
                ENTERING atomcount (s = nil)
                EXITING    atomcount (value: 0)
              EXITING atomcount (value: 1)
            EXITING atomcount (value: 2)
          EXITING    atomcount (value: 3)
          ENTERING atomcount (s = nil)
          EXITING    atomcount (value: 0)
        EXITING atomcount (value: 3)
      EXITING atomcount (value: 4)
    EXITING    atomcount (value: 7)
    ENTERING atomcount
        (s = (ab ((cb ca ba) cb (ac ab cb))))
      ENTERING atomcount (s = ab)
      EXITING    atomcount (value: 1)
      ENTERING atomcount
          (s = (((cb ca ba) cb (ac ab cb))))
        ENTERING atomcount (s = ((cb ca ba) cb (ac ab cb)))
          ENTERING atomcount (s = (cb ca ba))
            ENTERING atomcount (s = cb)
            EXITING    atomcount (value: 1)
            ENTERING atomcount (s = (ca ba))
              ENTERING atomcount (s = ca)
              EXITING    atomcount (value: 1)
              ENTERING atomcount (s = (ba))
                ENTERING atomcount (s = ba)
                EXITING    atomcount (value: 1)
                ENTERING atomcount (s = nil)
                EXITING    atomcount (value: 0)
              EXITING atomcount (value: 1)
            EXITING atomcount (value: 2)
          EXITING    atomcount (value: 3)
          ENTERING atomcount (s = (cb (ac ab cb)))
            ENTERING atomcount (s = cb)
            EXITING    atomcount (value: 1)
            ENTERING atomcount (s = ((ac ab cb)))
              ENTERING atomcount (s = (ac ab cb))
                ENTERING atomcount (s = ac)
                EXITING    atomcount (value: 1)
                ENTERING atomcount (s = (ab cb))
                  ENTERING atomcount (s = ab)
                  EXITING    atomcount (value: 1)
                  ENTERING atomcount (s = (cb))
                    ENTERING atomcount (s = cb)
                    EXITING    atomcount (value: 1)
                    ENTERING atomcount (s = nil)
                    EXITING    atomcount (value: 0)
                  EXITING atomcount (value: 1)
                EXITING atomcount (value: 2)
              EXITING    atomcount (value: 3)
              ENTERING atomcount (s = nil)
              EXITING    atomcount (value: 0)
            EXITING atomcount (value: 3)
          EXITING atomcount (value: 4)
        EXITING    atomcount (value: 7)
        ENTERING atomcount (s = nil)
        EXITING    atomcount (value: 0)
      EXITING atomcount (value: 7)
    EXITING atomcount (value: 8)
  EXITING atomcount (value: 15)
15
—>
```

The Lisp function "atomcount" is traced as it is running on the list called "brahma"

simple descent into a smooth-walled canyon and then a simple climb back up the other side; this recursion looks like a descent into a craggier canyon, where on your way up and down each wall you encounter various "subcanyons" that you must treat in the same way—and who knows how many levels of such structure you will be called on to deal with in your exploration? The shape described by a structure that goes on indefinitely like that has been named a "fractal" by Benoit Mandelbrot.

Notice in this recursion that we have more than one type of embryonic case (the "null" case and the "atom" case) and more than one way of descending toward the embryonic case (by way of both car and cdr). Thus our Big Questions can be revised a bit further:

1*a*. Is there just one embryonic case, or are there several cases, or is there even an infinite class of them?

1*b*. How can you know when you have reached an embryonic case?

1*c*. What are the answers to the various embryonic cases?

2*a*. From a typical case is there exactly one way to step toward an embryonic case, or are there various possibilities?

2*b*. From a typical case how do you determine which of the various routes toward an embryonic case to take?

2*c*. How do you build this case's answer out of the "magically given" answers to one or more simpler cases?

One of the most elegant recursions I know of originates with the famous disk-moving puzzle known variously as "Lucas's Tower," the "Tower of Hanoi" and the "Tower of Brahma." Apparently it was originated by the French mathematician Édouard Lucas in the 19th century. The legend that is popularly attached to the puzzle goes like this:

In the great temple of Brahma in Benares, on a brass plate under the dome that marks the center of the world, there are 64 disks of pure gold that the priests carry one at a time between three diamond needles according to Brahma's immutable law: No disk may be placed on a smaller disk. In the beginning of the world all 64 disks formed the Tower of Brahma on one needle. Now, however, the process of transfer of the tower from one needle to another is in midcourse. When the last disk is finally in place, once again forming the Tower of Brahma but on a different needle, then will come the end of the world and all will turn to dust.

A picture of the puzzle is shown below; I have labeled the three needles "a", "b" and "c".

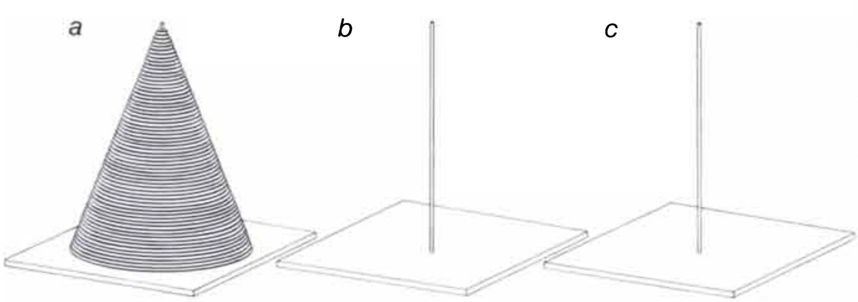
If you work at it, you certainly can discover the systematic method the priests must follow to get the disks from needle "a" to needle "b". For only three disks, for instance, it is not difficult to write down the order in which the moves go:

ab ac bc ab ca cb ab

Here the Lisp atom "ab" represents a jump from needle "a" to needle "b". There is a structure to what is going on, however, that is not revealed by a mere listing of such atoms. It is better revealed if one groups the atoms as follows:

ab ac bc   ab   ca cb ab

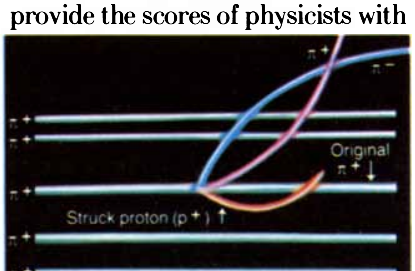
The first group of three accomplishes a transfer of a 2-tower from needle "a" to needle "c", thereby freeing up the largest disk. Then the middle move, "ab", picks up that big, heavy disk and carries it over from needle "a" to needle "b". The final group of three is much like the initial group in that it transfers the 2-tower back from needle "c" to needle "b". Thus the solution to moving three disks depends on being able to move two. Similarly, the solution to moving 64 disks depends on being able to move 63. Enough said? Now try to write a Lisp function that will give you a solution to the Tower of Brahma for n disks. (You may prefer to label the three needles with digits rather than letters, so that moves are represented by two-digit numbers such as 12.) I shall present the solution next month—unless, of course, the dedicated priests, working by day and by night to bring about the end of the world, should chance to reach their cherished goal before then.



The Tower of Brahma puzzle. The 64 disks on "a" must be placed in the same order on "b"

of data from these experiments, which may run for weeks or months.

As Charles Dickens, director of SLAC computing services, says, "Our 3081 moves data in and out of its central processor fast enough to provide the scores of physicists with



A significant event at SLAC. Positive pion (π^+) strikes proton (p^+) creating two new pions, one negative and one positive. Four other pions pass through with no contact.

immediate access to experimental results."

They can think about the problem, and not worry about the computer, with microcode-assisted VM/CMS. With this flexible IBM software and the 3081's reliability, physicists can, with confidence, steal a look for a significant event during the experiment.

Dickens explains, "Before, people might say, 'If I'd only known what was happening, I'd have done things differently.' Now they see the results as they happen."

The 3081 supports 450 terminals for interactive computing under VM/CMS. It performs well in online environments because a new dyadic design allows its two separate processors to work concurrently on a wide range of applications. Its high reliability is due, in part, to an advanced thermal conduction module (TCM).

IBM offers engineering and scientific users extensive support: consultants, education and access to professionals. Tap into our many years of experience.

For an informative brochure, write Dr. Jack W. Hugus, IBM Engineering and Scientific Marketing, 1133 Westchester Avenue, White Plains, NY 10604.

IBM®

BOOKS

Eskimos in a modern milieu, bats, Gothic cathedrals, fractals, the heyday of steam

by Philip Morrison

THE LAST KINGS OF THULE, by Jean Malaurie. Translated from the French by Adrienne Foulke. E. P. Dutton, Inc. ($25.75). The Greenlander Knud Rasmussen, poet, ethnographer, explorer, the man "whose laughter went before him," holds sway over this book. His fancy named the place, his example led the author, his brooding face looks out of the frontispiece, and the small white house where once he lived still stands on the foreshore at Thule Air Force Base, before the huge phased-array radar and the big runways. But the kings of Thule, the polar Eskimos in archetype, the hunting people of that oasis along the northwestern corner of icy Greenland, have left forever. What they were like, how they came to their reign, how they live now and what the future may hold for them form the learned, anecdotal and heated chapters of this book of loving witness.

A remarkable history slowly emerges from behind the details of daily life and the dramatic personal encounters that crowd these pages. The place is a kind of second Bering. The complex conformation of the ocean channels has meant rich sea life and delayed freezing for so high a latitude. There is no easier place to cross from Canada to Greenland. The hunters found their way to the island continent at least as early as the 12th century. They knew land and sea, they took seal and walrus from kayak, snared the great flocks of guillemots, fished for fat salmon. The coasts are narrow in front of the ice cliffs; the people were only a few hundred in all the region, a couple of hundred miles of infolding shores. Then the climate changed. Between 1600 and 1800 a little ice age arrived, and the sea mammals became few. The birds were the mainstay, and the people suffered. In 1818 outsiders found them. Sir John Ross described them: without kayaks, without knowledge of how to take salmon or use bow and arrow, the old skills lost. Ross showed them wood for the first time. They had iron, some from meteorites lying on the coast, a little from Viking trade. From then on whaling ships came by occasionally, with wood and iron.

In about 1860 the sorcerer Qillarsuaq, who lived among the Canadian islands more than 500 miles to the south, was "seized by the prescient certainty" that much farther north were living hunters still unknown. He persuaded his own people to join him for an indescribably difficult journey into the unknown. Had his "white hair not been haloed by a nimbus of fire every night" who would have stayed the course for seven seasons? In the end they found lots of men and women and their sledges. The Canadian Eskimos settled in the new place, reviving the arts of kayak and bow.

Now came the polar searchers, Peary and Cook. Peary came every five years or so, engaged half of the hunters to staff his journeys, took back the great meteorites to New York, traded for walrus ivory and fox furs, on terms of trade most unfavorable to the hunters. After 20 years Peary was a hero to the Inuit, even though he had never dealt fairly with them. In his decisive way, however, he had brought them rifles and ammunition. In 1951 Malaurie met the two sons whom Peary and his long-time companion, the black seaman Matthew Henson, had left behind. There are persuasive photographs of these two hunters, men ignored in the U.S.

After Peary's somewhat uncertain claim on the Pole he returned no more. Now it was Rasmussen's time. He established the trading post at Thule—so he styled the old Inuit village of Uummannaq, the seal's heart—to supply the hunters, taking pay in the beautiful blue-fox fur they brought in proudly. The Danish administration did not want to extend its trading and service network that far north. Until Rasmussen died in 1933 he was in effect the government of the people, and the proceeds of their trade supported his seven Thule expeditions that laid the foundations of Eskimo history. Some 300 polar Eskimos financed the finest program of studies of their own past, over a front of 9,000 miles. The hunters revere him still; he was one of them, lofty of purpose, respectful of dignity, entirely trustworthy.

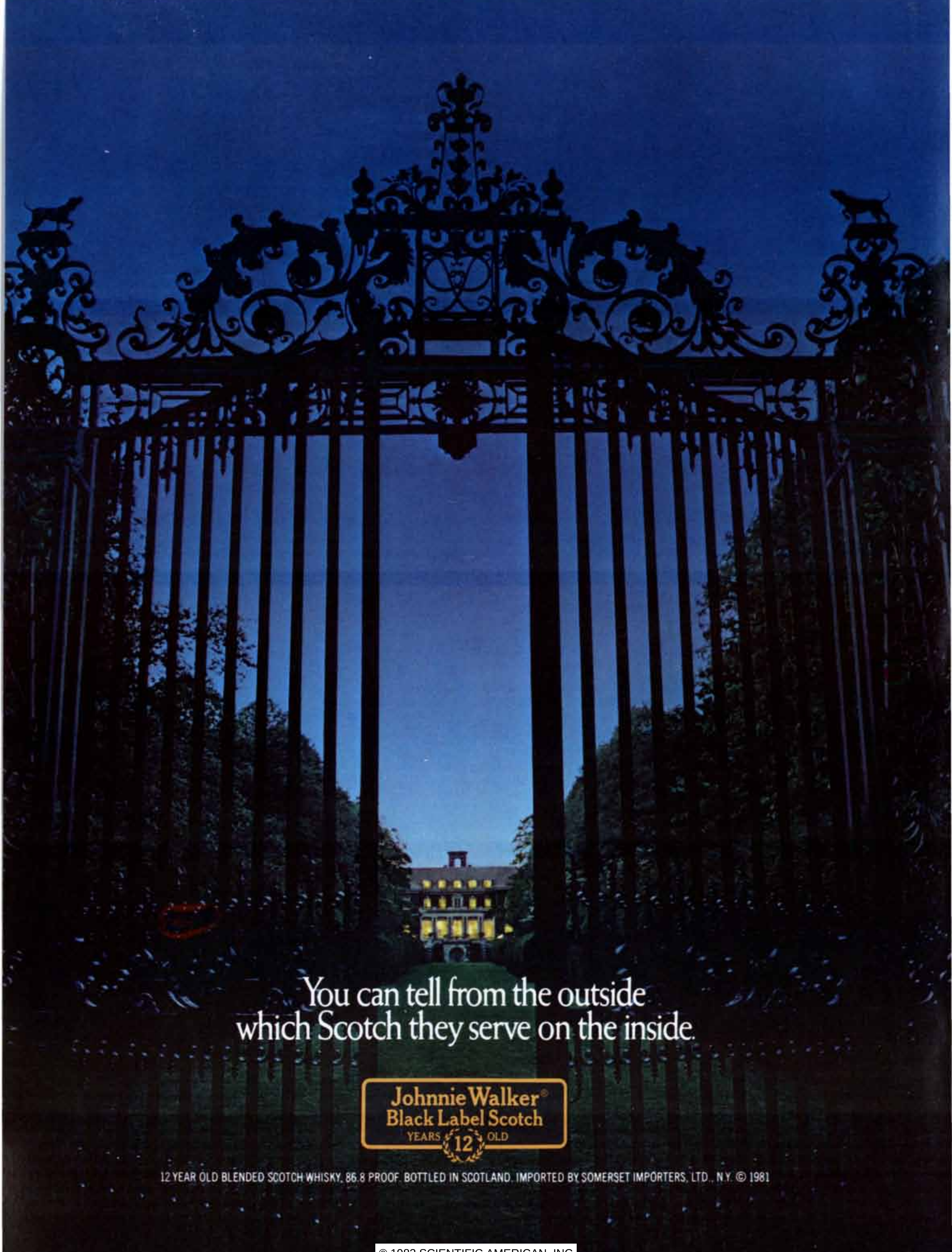
It was in 1950 that Malaurie came to the people of Thule, straight from the Sahara, where he had been mapping with Tuareg guides. He was a geomorphologist with a strong desire to know the people of the ends of the earth. His project was a careful aneroid survey of the beaches of the region, as a contribution to the understanding of the relaxation of glacial stresses. The Inuit were sophisticated enough by 1950, although they hunted still for a living. They had never, however, heard French. They liked the tongue, the fables of La Fontaine and the robust and competent young scientist with such a strong sense of character. He published his survey maps, which had cost a long, hard journey with a party of hunters and their families, but geology was now behind him. Since then he has been a student, companion and filmmaker for Eskimos around the circumpolar world.

As Malaurie came to Thule on his homeward way he encountered an event that cut across the history of the region with the decisiveness of a slip along some great fault. In the summer of 1951 the lonely bay at Thule became host to an unprecedented flotilla of ships; the shore was forested with construction cranes; 20,000 men and their machines arrived to work for the U.S. Strategic Air Command. By 1954 the government of Denmark had built the hunters a new village to the north, where they live today. The SAC base holds 3,000 airmen still. The blue foxes live off the generous Air Force dump. The guards shoot them as vermin now, the pelts of no interest.

Elsewhere the foxes are scarcer. And the kings of Thule? The hunters stay home; they feed on tea and sugar, with smokes and drinks in plenty. At the age of 50 those who were hunters in 1951 can no longer chew meat. Danish welfare has replaced the arduous hunt.

It need not be. Indeed, hope is present. By 1980 it looked like "Eskimo Gaullism," Inuit power. Men of Thule came to the all-Greenland meeting in the south proudly wearing the rare bearskin pants. The choice is not the isolated past of pride and terror, the kings vassal to the icy wind, infants sacrificed in dire need. Pay a fair price for the lovely furs. If the fox range must be extended, haul the hunters out by plane. They can return on their own, hunting as they go, the old skills augmented by the rifles, radios and compasses they can use. Support a strong emergency patrol system, the hunters out in the coldest times, maintaining a search-and-rescue network that might someday save the lives of air passengers downed along their great-circle route between two warmer climes. It may be that some form of husbandry for foxes and bears and seals can be developed, intermediate between the open hunt and the mink ranch. It will all cost the taxpayers of Copenhagen less than the welfare state, and it can allow those Inuit who do not treasure continuity to seek a newer way.

It might also preserve the kingship,


You can tell from the outside
which Scotch they serve on the inside.
Johnnie Walker
Black Label Scotch
YEARS 12 OLD
12 YEAR OLD BLENDED SCOTCH WHISKY, 86.8 PROOF. BOTTLED IN SCOTLAND. IMPORTED BY SOMERSET IMPORTERS, LTD., N.Y. © 1981

the royal view that "we are the strongest" or the phrase that "for an Eskimo it's easy." Like the kings of other latitudes, these nomads of Thule are not always admirable. They are often vain, mercurial, harsh. Their consorts, the woman tied to the hunter both by a strong sexuality and by the most urgent necessities of mutual survival, were granted dignity and status, but they lived too hard a life, worn out long before their men. We cannot preserve all the past, but this human past is symbolic for us all; did not our species embark on its strange, swift adventure once the edge of the ice receded?

ECOLOGY OF BATS, edited by Thomas H. Kunz. Plenum Press ($49.50). Although they are furry, breast-fed kin of ours, the bats have a bad press, whereas their feathered reptilian predecessors in the air catch all eyes. The reason is in the eye itself; bats are in a way the nocturnal equivalent of birds. They are successful aeronauts everywhere, diverse and abundant except at sea and toward the poles. Their world, however, is dark and colorless, so that we diurnal primates can watch it little.

This volume is a collection of 10 expert reviews of what bat watchers have lately come to know. The past 20 years have placed this order of nearly 1,000 species closer to its proper position in our understanding: it is the second order among mammals. (There are still more rodents than bats.) Bat studies are not easy; night work is the rule, with snooperscopes or telemetry receivers, with sonar translators or deep in caves filled with guano. In the laboratory too bats are taxing subjects. The book, whose level is that of the research survey in a set of well-written examples, serves two purposes for the general reader: it gives a clear account of its topic (if one heavy in taxonomic jargon) and it presents an exhibit of the detailed, mathematicized, ambitious (and still incomplete) state of field biology today.

True to its title, the collection tends to look away from the form and function of single bats. There is little on flight aerodynamics (perhaps too little) and almost nothing on the sonar system: the fierce mouths for pulsing shrieks and the labyrinthine directional ears. Roosts and fleas, reproduction, growth and survivorship, energetics and size relations, feeding strategies for both insects and fruit—these are the topics of comparison and conjecture. A motif that recurs, rather wistfully, is the desire to determine ecological niches from the form and ability of species. It is hard to see how it is that so many distinct species with such distinct pulses hunt so well the same mixed bag of bugs in flight. There is a lot going on we do not know.

Most of the larger Old World plant-visiting bats lack echolocation. The biggest have a wingspread close to six feet. They roost externally, often in the very trees where they feed in the warm tropical nights. Mousy Temperate Zone bats, gregarious in their caves by the 10 million, or widespread in the belfries, are insectivores, demon echo-locators of insects by dawn and dusk. Their enclosed roosts call for fancy navigational skills without light but reward them by keeping out the freezing chill of winter. Birds that in warm weather live on the insect hordes generally migrate a long way toward the Tropics once cold weather ends the buzzing life for the season. The bat counterparts migrate in time rather than space; they take the winter off not in Florida but in hibernation, a complex and well-studied state, energy-economic because the animal's temperature approaches the ambient one. This strategic choice is not yet understood.

In flying at night bats avoid the fiercest birds of prey. It has lately been recognized that owls, the silent night hunters, take an important toll of bats. The deeds have been witnessed from blinds with the infrared image amplifier, and the search of owl pellets for bat bones has confirmed the indictment. In the rich study of bat and light rhythms reported here one striking graph shows the measured activity of two bat species and two owl species as a function of light levels. There is overlap, because bats need enough night hours of work to make a living. But the peaks are sharply displaced: bats prefer to work at the light level of the dark of the moon, owls at that of the full moon. During a recent lunar eclipse the activity of bats showed an unexpected small-hours peak.

Owls do not hunt bats by listening, as they hunt mice in response to the scurry in dry leaves. In order to capture a fast-moving, darting bat the owl must track it at longer distance and strike swiftly. Deep darkness is the shield of the echolocating bat, even though the dark cannot hide the bat's insect targets from the sonic pulse or confuse its quick flight into the twisted crevices of home. In this connection the bat mother in the metropolis of the bat cave does not seek out her own infant. She nurses the first hungry kit she can find; her extended family can number in the millions.

Bats seem to live long for their size, both actuarially and physiologically. No fossils disclose any prebat stage before bathood. Perhaps the bats originated with a tiny tropical omnivore of the trees, first giving rise to a small flying pursuer of insects, then radiating into larger forms able to live on fruit, nectar or meat. The animals later developed a complex system of thermoregulation, and they could reproduce somewhat independently of external conditions. The hibernators of our time are to be seen as late-arising specialists of the seasonal world outside the Tropics. The cave is fully home now to the little bats, as much later it became a refuge and a home to our own lineage.

EXPERIMENTS IN GOTHIC STRUCTURE, by Robert Mark. The MIT Press ($15). The tall pages of this not very large book are fitting for the beautiful tall forms that stand on many of them: elegant sections of cathedral naves in sharp line, photographs of the entire masonry fabric of those great buildings, sections of plastic models mapped in the rainbow colors of stress interferometry. The argument too is elegant; this Princeton engineer on the site and in the model laboratory has probed for understanding behind the swift progress of the master builders of medieval Europe. Those designers and the teams they led, not serfs but "free—nay—unionized" masons and carpenters, learned in two centuries to capture ever more soaring space in their common stone.

The issue is not a new one. For a century the architect-restorers (Viollet-le-Duc is the most renowned of them) and the historians have joined over it. Structure, however, is not simple. Only recently have we been able to examine in some detail examples geometrically complicated enough to be relevant. Insights alone give equivocal answers to many of the questions the structures raise; Mark's models allow judgments that archaeology and argument cannot by themselves decide.

Mortar-jointed masonry gave rise to progress: it made possible the great towers and aisles. Even the highest belfry does not out-top the solid Pyramid of Cheops; stone on this scale does not fail under compressive load alone. But simple stability, each block in balance on the block below, does not change with scale. A model of toy blocks could serve for the lofty cathedral of Reims; indeed, such models may have been used. More important, the builder's experience would stand him well as his ambitions and structures grew. Masonry supports little tension; if the forces impose tensile stress, the masonry will soon fail visibly. Localized areas of tension cracks give a clear sign of incipient instability.

So armed, the builders could build with understanding, although surely they had no mechanical theories. The flying buttresses in building after building show their virtues in model tests, as for a long time the analysts have expected. Nobody "of sound mind" could take away the fliers from Reims, just as no one would now remove them from Nervi's ferroconcrete shells in Rome.

As the churches towered higher, the lateral wind loads became more important. The models disclose the entry of intolerable although localized tensile stress under realistic if unusual winds. From this we can infer the old builder's care in watching for such signs, and his modification of the design to cure it. The ornate pinnacles that weigh down the high piers are often—not always—effective measures against such failure. Moreover, it is clear that a cathedral is no fixed structure; constant repair today is evidence of the passing seasons in these economical and open designs. Beauvais, at 50 meters the highest of all towers, collapsed twice, once before 1300 and again in about 1570 (when it had been audaciously rebuilt with the tower 150 meters tall). The models suggest (the inference cannot be final because the structure is not well recorded) that the wind loads induced cracking within a poorly visible set of intermediate pier buttresses.

The art continued; Palma Cathedral is much more spacious than Reims. Its nave is more than twice as high; by one measure it is twice as slender. Heavy external pier buttresses, artfully arranged to house side chapels, have opened the interior safely without loss of the visual harmonies. Vaulting is still more difficult to analyze without models; even the ribbed barrels of the St. Louis airport, let alone the wonderful ceilings of the French High Gothic, do not disclose their structural necessities in visible form.

When Professor Mark was a college engineering student, he incautiously carried a volume on Picasso to help him wait out an early interview. A sympathetic engineer observed: "Young man, never bring a book like *that* when you look for a job in engineering." This book shows his independence of mind. He has not, however, forgotten the broader demands of unity: both the beauty of the volume and the careful explanatory apparatus—not one equation but plenty of serious mechanics—fit it alike for humanly curious readers with engineering savvy and for those who deal freely only with word and form but remain open to the support of quantitative reason.

THE FRACTAL GEOMETRY OF NATURE, by Benoit B. Mandelbrot. W. H. Freeman and Company ($32.50). In 1975 there appeared a small but dense and delightful volume in French around the neologism "fractal," the invention of this gifted, playful and tirelessly imaginative author. That work, "macédoine de livre," was reviewed in these columns with the enthusiasm it deserved. It was promptly enough served up in English, the key word still visible, the book fuller, much more visual, extended in scope and form. Here it is once again, the new edition straightforward of title, bearing a set of color plates that exhibit the virtuosity of today's programmers, the mixture as fragrant and fresh as before.

The new form, perhaps increased a third in size, is ready for lucky readers who have not yet found their way to fractals, no less than for those who simply want more. Meanwhile the topic has become all but trendy. Strange attractors and renormalization groups on a lattice, topics that blaze in the physics colloquiums these days, are here, full of evocative content albeit incomplete. The subject matter of the older editions is not neglected: the simplest idea behind the concept leads still to the understanding that the coastline of Maine has a length dependent on the inquirer. A ruler laid on the map from Portland to Calais gives a minimum answer; a line drawn through the indentations of bays and headlands on a highway map gives a longer one; the ruler put down centimeter by centimeter at real tidewater gives a much longer one, including every cobble and pebble.

The point is made quantitatively and is related to the notion of dimension. Such an exquisitely tortuous curve falls between the dimensionality of a line and that of a plane, reasonably enough. Hence the idea and the word "fractal." Of course, chance must enter, although rich deterministic examples are found in the dizzying idealized constructions of the long-known theory of real variables. Let the matter rest there; those who wish can find their way to the subject easily.

The book is no text, no set of definitions, no series of applications. It is a flowing source, as once the wonderful *On Growth and Form* by D'Arcy Wentworth Thompson came to us all. It too belongs on the small shelf of books that disclose the forms of nature. Thompson, classical in approach, centers on the simple forms hidden in the complex web of life. Mandelbrot, resting squarely on the mathematical moderns—Cantor, Koch, Hausdorff, Paul Lévy—opens instead a path to the simplicity of plenitude. His heroic motifs are not spheres or spirals but intricate ideals of islands and reefs, stars and curdled milk, turbulent spray and pink noise, the abstract forms of nature's generative profligacy. Naturally enough, the book itself is less lucid, but it is unifying, recursively fascinating, agreeably personal without posturing and buoyant with scientific hope.

THE POWER OF STEAM: AN ILLUSTRATED HISTORY OF THE WORLD'S STEAM AGE, by Asa Briggs. University of Chicago Press ($22.50). Frontispiece and jacket bear the title of the volume in an artful style: the legend is a cast brass nameplate, evocative of the Steam Age less only than the hiss, the throb and the scent of oily rags. From a photograph of a Newcomen engine in place (1855), past the Exhibition of 1851, the wire mill of Sharp and Brown, a Victorian smokestack topped with Norman crenelations, the *Great Eastern* in New York to turbines of our day the illustrations present an overview both canonical and fresh. Sooty miners and spruce bellhops, sheet music and racing paddle-wheelers, the view is a wide one.

The book seeks to "bridge the gulf" between the pictorial albums and the austere volumes of economic history. Lord Briggs is a distinguished cultural historian of the period; his text is supple and persuasive, although it is more a fragrant broth than a savory hot pot. The beginnings are particularly well put. Steam power, almost a century before James Watt, served to pump the big collieries dry; England was a smoky industrial country before the Industrial Revolution. Textile factories, particularly for the fulling of woolen cloth, were users of considerable power long before Watt, driven by water and even by horses. Watt rated his engines, after all, in horse power, a comparison quite familiar to his customers. Karl Marx said it well: "The steam engine itself... did not give rise to any industrial revolution. It was... the invention of machines that made a revolution in the form of steam engines necessary."

Steam power, both the stationary engines in every shop and mill and the locomotive engines by land and sea, gave a kind of unity to the rise of industry and to its diffusion worldwide. Lewis Mumford spoke of "carboniferous capitalism." For Briggs there arose rather a "gospel of steam," a vision of global progress through pistons. What happened was worldwide, but it was at once both richer and crueler than the poets and the reformers could have foreseen.

The book ends in our time of internal combustion and electronics. Locomotives and tractors are diesels these days; the reciprocating steam engine has become a piece of archaeology. The text does make the point that spinning turbines under high-pressure steam still dominate the big power plants, but it does not enough emphasize that the bulk of the electric power of our world is still the work of steam turning steel shafts. Indeed, the last chapter is frank nostalgia: vintage steam. In Britain there are some 200 sites where pistons under steam can be seen alive and at work today or in a specialized museum setting. About 30 sites are named in the U.S., and others are specified worldwide.

Asa Briggs is chancellor of the Open University. Although he has read Henry Adams, he seems to cavil at the late Lord Snow's plausible remark that the laws of thermodynamics might be held as material even a literary person ought to know. Laws narrowly "drawn up to explain the workings of steam engines" in fact hold universally, and cast the coldest of eyes on life, on death. The two cultures still walk two paths. This book was conceived and assembled by a clever British editorial firm. The absence of meaningful sources for the many fine illustrations is the only deficiency of this attractive and readable volume.

Established 1845 SCIENTIFIC AMERICAN March 1983 Volume 248 Number 3

The Law of the Sea

In December 119 nations signed the Convention on the Law of the Sea. The U.S. was not among them, but development programs under the new law will nonetheless be proceeding

by Elisabeth Mann Borgese

On December 6 of last year the representatives of 119 nations, gathered at Montego Bay in Jamaica, signed the United Nations Convention on the Law of the Sea. The ceremony culminated 15 years of labor by the largest and longest of all international conferences. The Third United Nations Conference on the Law of the Sea had engaged as many as 3,000 delegates in 11 long sessions totaling 585 days. Beginning in the chaos and conflict of contending political and economic interests, working through draft after draft, the Conference succeeded in writing a "constitution for the oceans" in 320 articles (organized in 17 main parts) and nine technical annexes.

If the gist of this enormous work can be stated in one sentence, the Convention replaces the traditional laissez faire system of freedom of the seas with an emerging system of management. In place of the two-dimensional boundless sea the Convention deals with a finite three-dimensional resource, its depths of as much economic interest as its surface. It puts 40 percent of the ocean and its bottom adjacent to the coasts of the continents and islands under the management of the states in possession of those coasts. It reserves the other 60 percent of the surface and the water below for the traditional freedom of the seas, but it deeds the wealth of the ocean floor, 42 percent of the earth's surface, to the Common Heritage of Mankind. The Convention places that heritage under the management of an entirely novel International Seabed Authority, which has the capacity to generate income, the power of taxation and a kind of eminent domain over ocean-exploiting technology.

Out of opposition to the Seabed Authority, it is true, the U.S., the U.S.S.R. and 15 other major industrial nations withheld their signatures from the Convention. These nations did, however, sign the Final Act of the Conference. (Only one nation, Turkey, signed neither.) The nations that signed the Final Act but not the Convention are entitled to attend future meetings of the agencies created by the Convention as nonparticipating observers. The signing of the Convention gives it the force of law, subject to ratification by at least 60 nations. When the Convention has that ratification, it can be expected that the nonsigners will sign as full participants.

The Convention provides a comprehensive global framework for the protection of the marine environment, a new regime for marine scientific research and a comprehensive system for settling disputes. It ensures freedom of navigation and free passage through straits used for international navigation, invoking the new legal concept of "transit passage," a right that cannot be suspended under any circumstances. It updates and codifies traditional law. It defines rights as well as responsibilities for states and international organizations in establishing and enhancing a new order in the world's seas and oceans.

The making of the new Law of the Sea engages all the issues the world community has to face in our age: food and energy, the arms race, communications and the "information revolution," trade, commodity policy, resource management and conservation, regional integration and the movements in science and technology that underlie many of these. It is as though all the issues were flowing together in the world ocean and the ocean had now become our laboratory for the building of a new world order. That order, we may hope, will prove more rational, more humane and more responsive to the real needs of the world than the old order that is disintegrating in hunger and violence.

Until 1973, when the Third United Nations Conference on the Law of the Sea began, the law of the sea had always been made by one or two or half a dozen seafaring states. Through the centuries, or the millenniums, the old-world seas came under the shipping code of Hammurabi and later the Rhodian sea law. The Mediterranean, the cradle of Occidental civilization, became the *mare nostrum* of Rome. In 1493 Spain and Portugal divided between them the world ocean beyond the Strait of Gibraltar. For the maritime power of 17th-century Britain, John Selden propounded the theory of *mare clausum*, the closed sea "owned" by the coastal states that would have the right, like the Spanish and Portuguese in their day, to control and exact tolls from navigators of other nations. It was against this doctrine that the Dutch jurist Hugo Grotius advanced the concept of *mare liberum*, the free ocean open to navigators from all nations, each respecting that right for others. In historical reality the freedom of the seas could be enjoyed only by those who could enforce it; throughout modern history they have been a handful of countries of western Europe and North America.

Two dramatic developments in human affairs, concurrent and often conflicting, compelled the calling of the Third United Nations Conference on the Law of the Sea. The first is the maritime dimension of the scientific-industrial revolution that is painfully transforming industrial societies and profoundly affecting their relations with preindustrial societies. The discovery of the phenomenon of sea-floor spreading has had the earthshaking, in the literal sense, consequence of installing plate tectonics as the central organizing idea in geology. That same discovery has far-reaching economic implications. Where the sea floor is rifting, at the axes of its spreading, metals selectively separated from the mantle of the earth well up to the surface. It is now understood that this process laid down the veins of nonferrous metals hitherto worked in the world's hard-rock mines. From the same closer acquaintance with the sea floor we know that it holds more oil and gas, in the continental shelves and in the benthic floor as well, than the traps in continental rock (much of it too, of course, once sea floor).

Meanwhile the rapid technological development of the fishing industry has undone the old notion that the ocean fishery is inexhaustible. The assumed infinity of the ocean has been, at the same time, overtaxed by pollution from urbanization, industrialization and rising population.

Thus the assumptions of the old law of the sea were eroded and the law began to crumble. The U.S. started the expansion of national claims to the ocean and its resources in 1945, with the Truman Proclamations on the Continental Shelf and the Extended Fisheries Zone. A scramble was on to carve up the world ocean, just as Africa had been carved up a century earlier.

While the Industrial Revolution was penetrating the ocean from the Northern Hemisphere the second great impulse that has shaped the new Law of the Sea was gathering force in the Southern Hemisphere. From the colonial empires assembled by the former maritime powers of Europe there came a procession of new sovereign nations. No more than 70 nations had participated in the First United Nations Conference on the Law of the Sea, called in 1958, or in the second one in 1960. By 1973, when the Third Conference convened, the membership of the UN had more than doubled. The majority in the General Assembly had passed to the poor nations of Asia, Africa and Latin America. They were determined to have their say in the deliberations of the Conference. Their call for a new international economic order was about to be sounded. The new Law of the Sea was to be part of it and even a model for it.

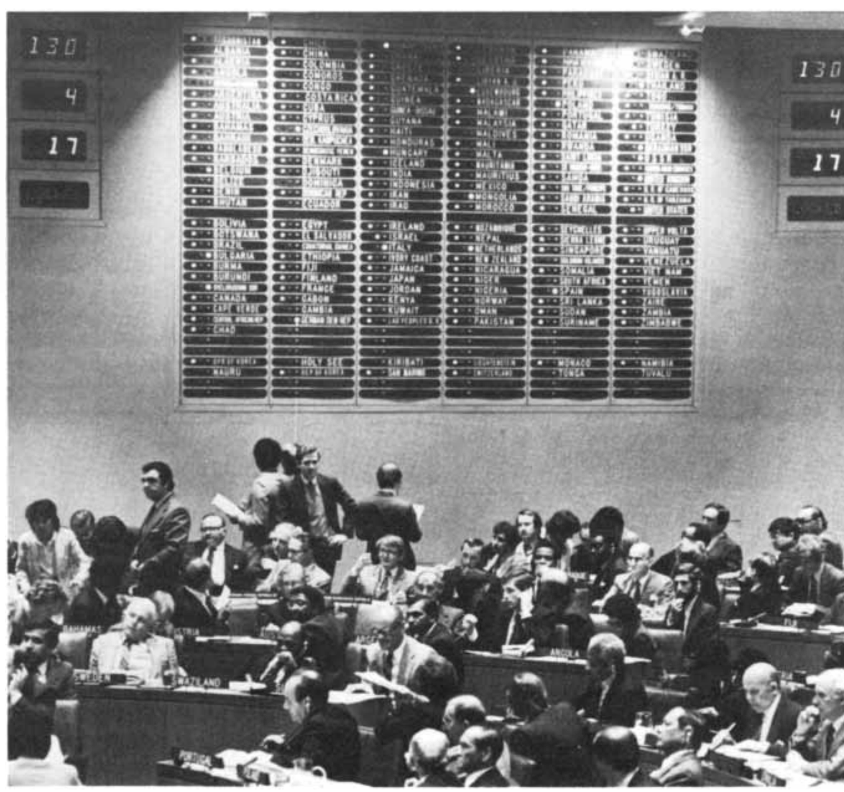
The Convention on the Law of the Sea clearly exhibits the shaping by the parallel yet often intersecting revolutions in technology and international relations. It was the technological revolution that pushed from the North the expansion of the national jurisdiction in ocean space. It was the revolution in international relations that pressed, from the South, for the establishment of new, strong international institutions. The two revolutions are equally in evidence in the massive record of the proceedings of the long conference. The Convention itself and the Conference record also reflect, however, the presence of a number of able, scholarly and statesmanlike personalities from both the North and the South: Arvid Pardo of the island state of Malta; Hamilton Shirley Amerasinghe of Sri Lanka, president of the Conference until his untimely death in 1981; Jens Evensen of Norway; Jorge Castañeda of Mexico; S. P. Jagota of India, and Louis Sohn of the U.S., to name only a few. The workings of their individual genius interacted with the collective pressures from the North and the South to steer the work of the Conference.

It was Pardo who proposed, in a historic address before the First Committee of the United Nations in 1967, that the wealth of the ocean floor should be declared the Common Heritage of Mankind. Neither the freedom of the sea nor the sovereignty then being claimed on the ocean by coastal states, he said, could ban the triple specter of pollution, of exhaustion of marine life and of international strife. Only international cooperation could preserve the peace and bounty of the oceans, and this might be furthered by the circumstance that vested interests were not yet as firmly entrenched in the oceans as they were on land.

A Committee on the Peaceful Uses of the Seabed was immediately appointed; a Declaration of Principles, incorporating Pardo's major thesis, was adopted in 1970, and the reservation of the seabed for peaceful purposes was incorporated in 1972 in the treaty that bans nuclear weapons from the seabed. In 1972 the seabed committee, still at work, called for the convening of the Third United Nations Conference on the Law of the Sea and began the preparation of its agenda.

The agenda was completed in 1973, when the work of the Conference on the Law of the Sea began. Organized in a General Committee, three main working committees and an intricate network of regional groups, interest groups, negotiating groups, plenaries, informal plenaries and a drafting committee working in six languages, it was the biggest and most complex conference in history.

The delimiting of ocean spaces—the traditional territorial sea and contiguous zone as well as the new extended economic and continental-shelf zones—was assigned to the powerful Second

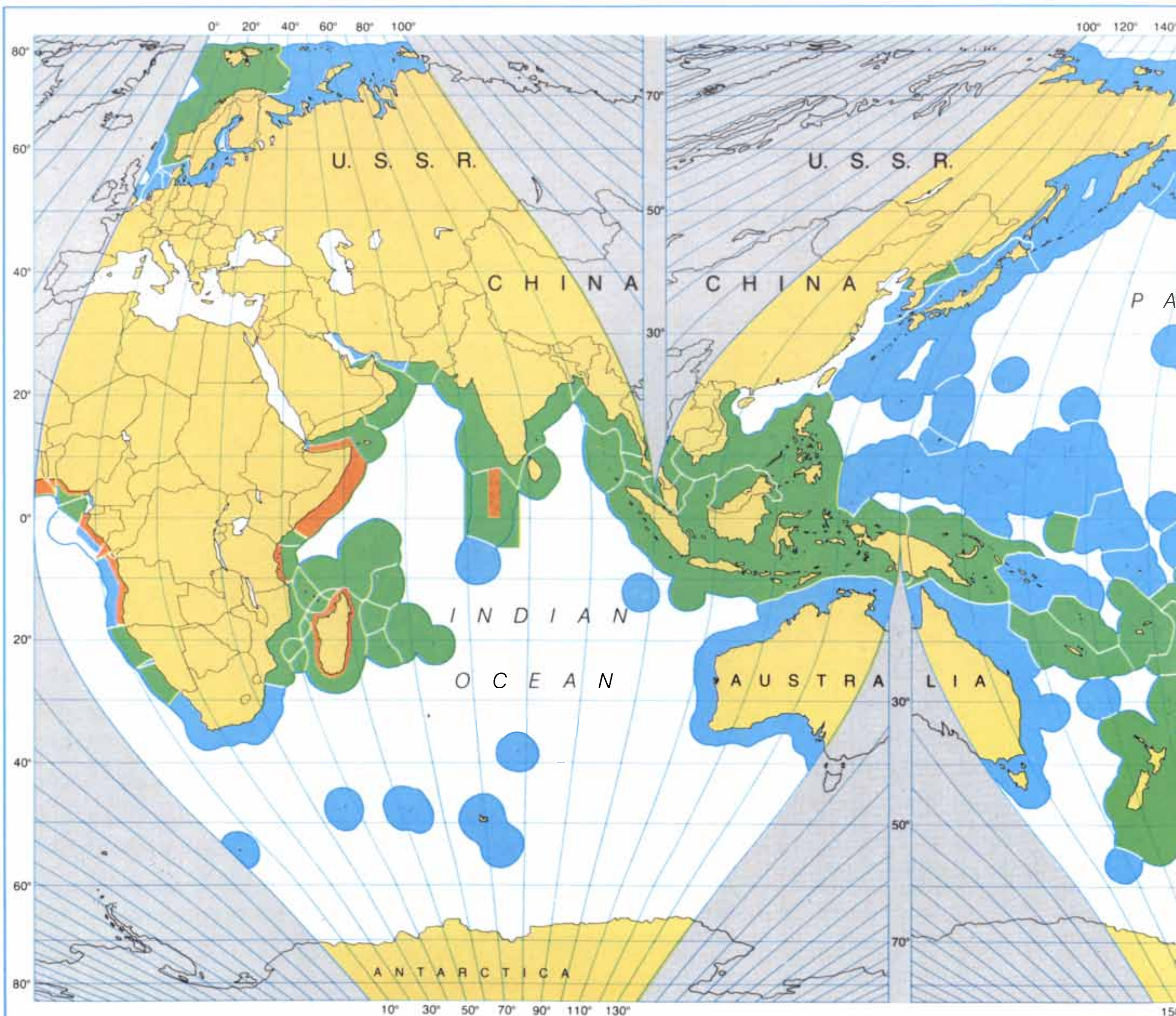


ADOPTION OF CONVENTION on the Law of the Sea was photographed at United Nations Headquarters in New York on April 30 of last year. Above the delegates is a tally board that records the vote: 130 nations in favor, four against and 17 abstaining. At Montego Bay in Jamaica on December 6, 119 of the nations signed the Convention and 143 signed the Final Act of the Conference on the Law of the Sea. Signing the Final Act but not the Convention entitles a nation to attend future meetings of the Preparatory Commission as a nonvoting observer.

Committee. Here the conflicting interests of three groups of nations had to be reconciled. The coastal states, of course, sought to extend the limits of their jurisdiction as far as possible and to secure maximum attainable rights in the zone under their jurisdiction. At the other end of the spectrum were the landlocked states, including the numerous landlocked African nations, the poorest of the poor. They could claim no zones and stood to lose what rights they might have enjoyed, particularly with regard to fishing, in the seas about to be closed to them. Somewhere in the middle were the maritime states, sharing some of the interests of the coastal states but depending, on the other hand, on the freedom of the seas to deploy their distant-water fleets and so sharing interests with the landlocked states. Each of these groups was a complex aggregation of developed and developing countries with both free-market and centrally planned economies.

There can be no doubt that the largest gainers from the expansion of national jurisdiction were the industrial states, particularly the U.S., Canada, Australia and South Africa, and also a few big developing countries, India, Brazil and Mexico among them. The articles of the Convention issuing from the Second Committee affirm the 12-mile limit of the territorial sea as well as a 12-mile contiguous zone. Beyond these limits they accord to the coastal states sovereign rights over the use of natural resources in the newly sanctioned 200-mile economic zone and in addition a bundle of other rights with respect to such activities as scientific research. Where the continental shelf extends beyond 200 miles the economic-zone rights with respect to resources on the shelf are extended with it, but not beyond 350 miles.

The coastal state does not hold territorial sovereignty in its economic zone, however, as it does over its land and ad-



MARITIME CLAIMS of nations are compiled in this map based on one compiled by F. Webster McBryde of Transemantics, Inc. The areas in blue are fishing zones, where exclusive fishing rights are claimed. The areas in green are economic zones, where not only fishing rights are claimed but also sovereign rights over all resources and jurisdiction over research and marine pollution. The areas in orange are "territorial seas," where full sovereign rights are claimed subject to the right of "innocent passage" (transit without military, economic

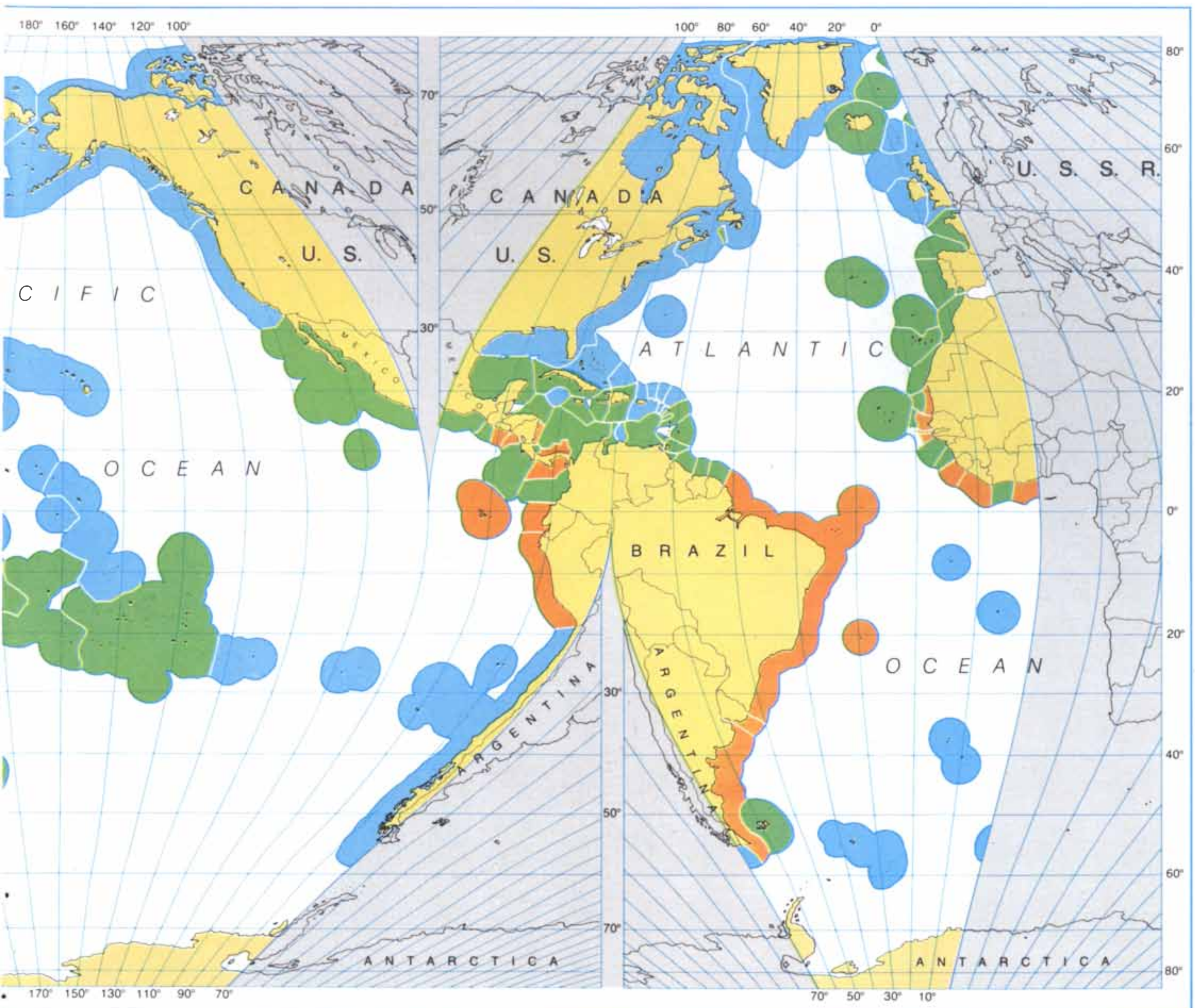
jacent territorial sea. This new kind of functional sovereignty may be regarded as the least malignant form of national aggrandizement. It is a new and dynamic concept in international law that undoubtedly will find other applications in the future.

The creation of the economic zone responds also to the need for management of the maritime environment. It provides an economic incentive for national action in this regard. It does not quite satisfy the need, however, since neither pollution nor fish respect national boundaries.

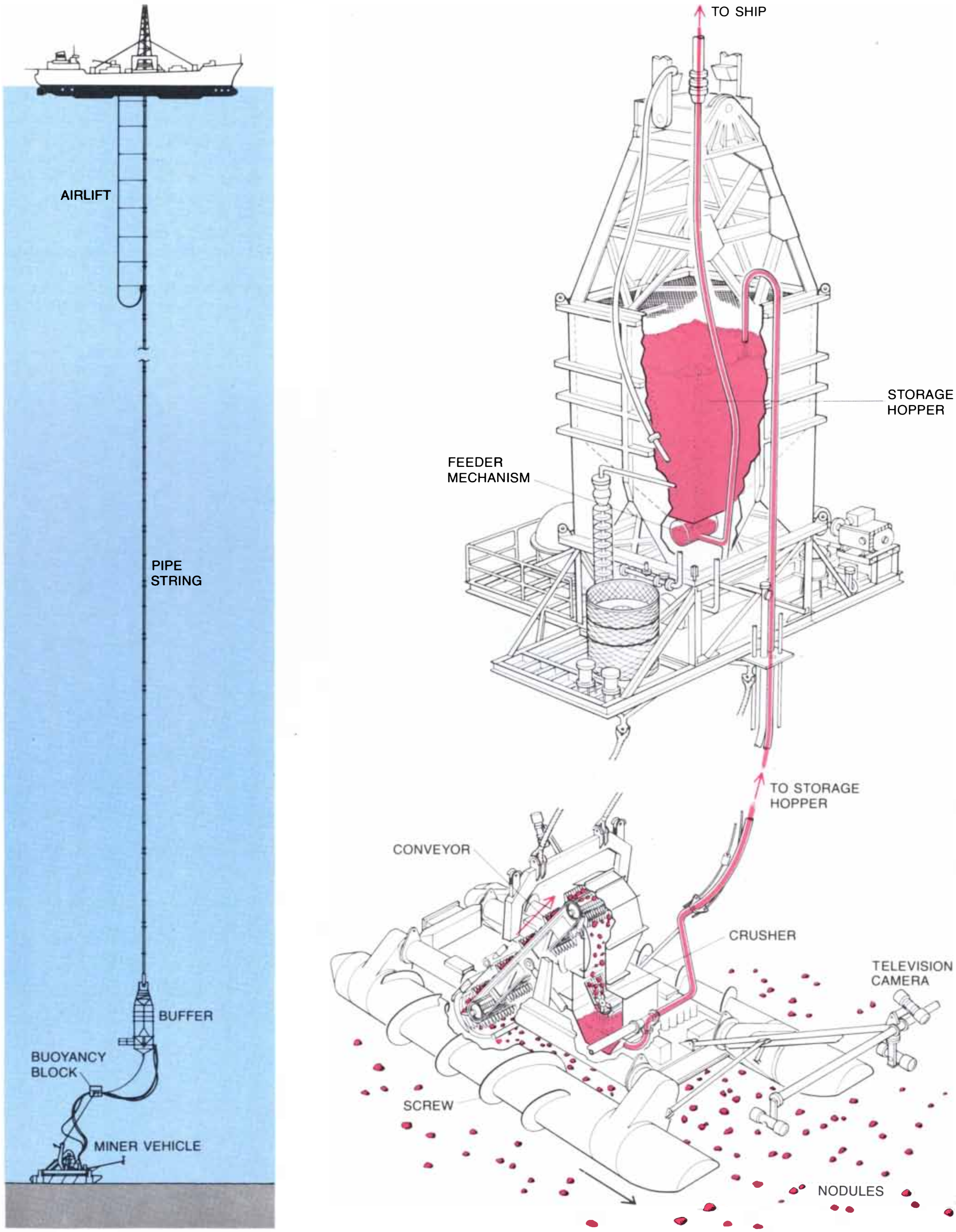
On the negative side the economic-zone concept has been called the greatest territorial grab in history. The expansion of national jurisdiction in the oceans surely increases inequality among nations. If it was hoped that the Conference would set clear and unambiguous boundaries to national jurisdiction, it must be said to have failed in this respect as well. The limits as defined are elastic, with loopholes for further claims at any time. It is an easy prediction that the discovery of any substantial resource anywhere in the world ocean will invite a claim from the nearest coastal or island state, in spite of the solemn declaration of the nonappropriability by any state or person (a corporation, for example) of the Common Heritage of Mankind "beyond the limits of national jurisdiction."

It can nonetheless be hoped that developments in the future will overcome the weaknesses and mitigate the injustices inherent in this part of the Convention. Most likely that will happen through progress in regional cooperation and the strengthening of international organizations. As will be seen, there is reason to look forward to some form of revenue sharing from the exploitation of the economic zone.

Political alignments were somewhat different in the First Committee, charged with setting up the international regime for the management of the Com-



or scientific significance). These claims, adding up to some 40 percent of the oceans, are recognized by the Convention on the Law of the Sea. The remaining 60 percent of the area is reserved for the traditional freedom of the seas, and the resources of that area of the ocean floor are deeded to the Common Heritage of Mankind. The projection is one designed by McBryde to avoid the distortion of both land and sea areas. The map can be obtained for $5 ($6.50 postpaid) from Transemantics, Inc., 1828 L Street NW, Washington, D.C. 20036.



MANGANESE NODULES, which are strewn over large areas of the ocean floor, could be harvested in a number of ways. The system shown here has been designed by Conrad G. Welling and his colleagues at the Ocean Minerals Company. The nodules typically contain 25 percent manganese, 1.5 percent nickel, 1.2 percent copper and .2 percent cobalt. In this system a miner vehicle (*bottom right*) is propelled across the bottom by two large screws. The nodules are picked up from the bottom by an endless-belt rake, crushed and pumped as a slurry to a "buffer" (*top right*) suspended by a pipe string from the surface ship. The slurry is then pumped to the ship.

mon Heritage of Mankind. Here the divergence of interests between the industrial nations of the North and the developing nations of the South was clearly the polarizing agent. The South held that the adoption of the Declaration of Principles by the General Assembly of the United Nations in 1970 had given the Common Heritage of Mankind the sanction of international law. In its view the task of the First Committee was to create a strong Seabed Authority with comprehensive functions and powers. The Authority would have an operational arm, the Enterprise, which would itself explore and exploit the Common Heritage on behalf of all mankind and could compel mining companies to sell it the technology needed for the task. Reacting to these proposals, the North proposed an Authority empowered to issue licenses and collect a small tax or royalty on the exploitation of the Common Heritage, which for the rest could be utilized by the North as it pleased.

The compromise that resolved this conflict was long and hard to come by. The Enterprise system conceived by the South was held to be an impractical proposition by the North, secure in its possession of the necessary capital and technology. The licensing agency advanced by the North was unacceptable to the majority of the nations convened in the Conference, who viewed it as a violation of the principle of the Common Heritage of Mankind. The compromise proposed by Henry Kissinger combined the impractical Enterprise with the unacceptable licensing agency in a "parallel system," a combination which, by a kind of witches' arithmetic, was to result in a solution both practical and acceptable! The Conference never quite recovered from it.

In the course of the tedious negotiations to spell out this compromise the text defining the parallel system became ever more abstruse and remote from reality. Playing on the fear of nonratification by the U.S. Senate, the American negotiator Leigh Ratiner succeeded in getting every administrative and financial detail spelled out in advance and engraved in articles that could not be altered for 25 years—and this for an industry not yet in existence and whose future was admittedly quite uncertain. The South meanwhile had second thoughts about the desirability of exploiting the Common Heritage at all at this time. Declining commodity prices made the resource-exporting developing countries wary of a new source of competition for their native ores. Against the opposition of the North (with the exception of Canada) these countries became more interested in limiting production from the seabed than in managing it.

The value of the Common Heritage of Mankind, particularly of the manganese nodules, has been discounted by other developments. Mexico and Chile have found rich fields of nodules in their economic zones; France has nodules in Polynesian waters, and the U.S. can assert claims offshore from the Hawaiian and other Pacific islands to hundreds of thousands of square miles of ocean floor likely to have commercially exploitable nodules. The nodules themselves have lost glamor to the deposits of polymetallic sulfides containing copper and other metals in much higher concentrations, first encountered at volcanic vents in the Pacific rift near the Galápagos Islands and off the coasts of Oregon and Washington. Similar deposits are certain to be found elsewhere along the 40,000-mile, world-girdling submarine rifted ridge.

Yet the creation of the International Seabed Authority by the consensus of practically the entire world community (in which the consent of the U.S. must be counted, because it was given at the Conference and will, I am sure, be given again) must be reckoned as a breakthrough in international relations. Here is an international institution unprecedentedly empowered to regulate and act on the basis of the new principle of the Common Heritage of Mankind. Here is a first attempt at a global production policy with due regard to conservation of the environment. Here is an opening to industrial cooperation between the North and the South based not on aid but on sharing. The European Economic Community was a dream of a few "federalists" 50 years ago. The International Seabed Authority, a utopian dream of 20 years ago, is now a fact of international law. Something has been moving.

The definition of a new regime for marine scientific research, for the protection and conservation of the marine environment and for encouraging the transfer of technology was the task of the Third Committee. Political polarization was less marked in the deliberations of this committee and it completed its work earlier than the others.

The freedom to do scientific research, one of the implied freedoms of the sea, is plainly constrained by the expansion of national jurisdictions into the ocean. It should be noted, however, that under the "consent regime" provided in the Convention a coastal state is normally expected to give its consent to nationals of other countries who want to do research in its waters. (U.S. scientists who have been worried about bureaucratic delays in getting coastal-state consent should now be concerned by their country's rejection of the Convention, which may prevent their even applying for such consent, much less getting it.) The consent regime should have the effect of encouraging bilateral cooperation in research enterprises, adjusting them to the interests of the coastal state as well as those of the investigators. More than any legal instrument in the past the Convention encourages the internationalization of marine research, time and time again making cooperation mandatory and calling on the "competent international organizations." Projects undertaken under the sponsorship of such organizations are deemed to have the consent of the coastal state, provided only that the coastal state has participated in, and not opposed, the project within the organization.

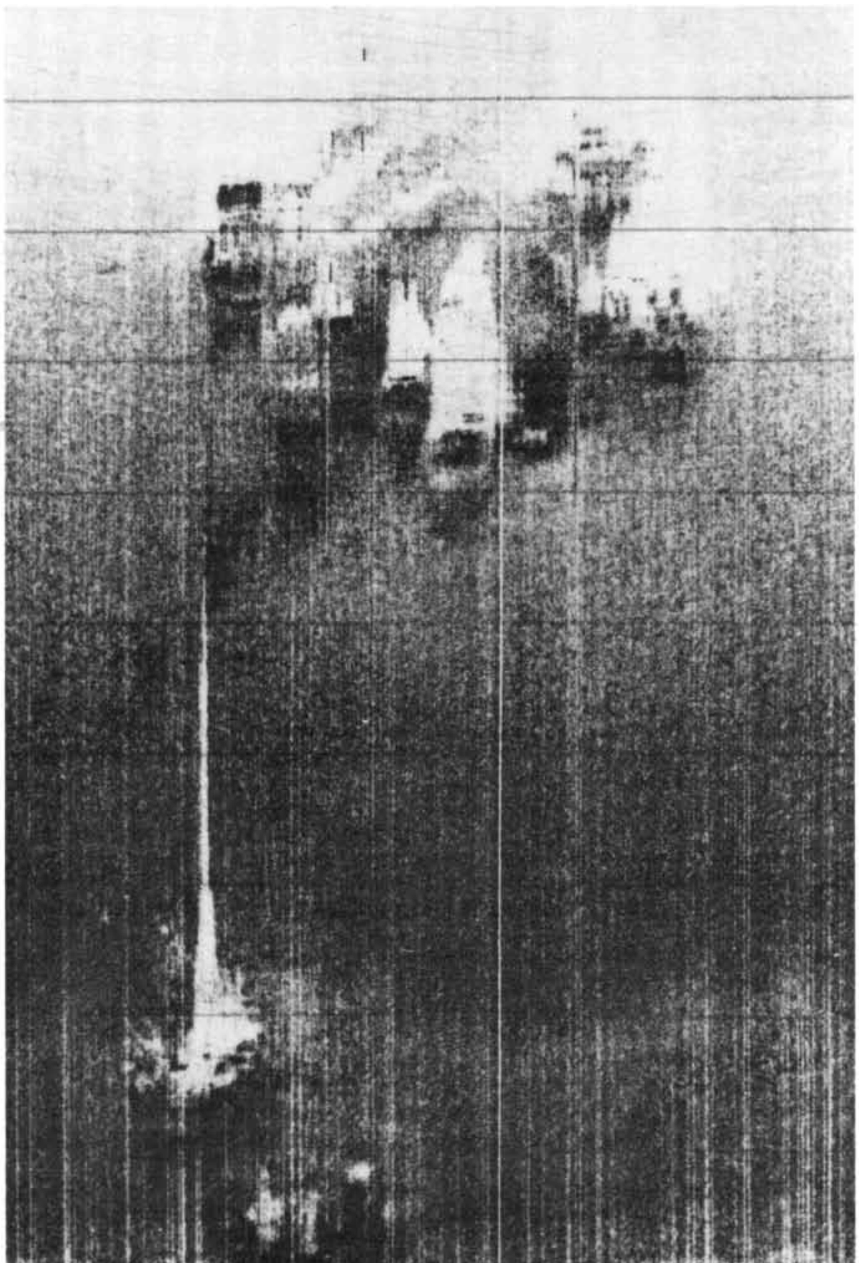
The measures for the protection and conservation of the marine environment provided in the Convention cover seaborne pollution, pollution from seabed mining and oil production and pollution from land-based sources and the atmosphere. This comprehensive framework for an international code of law even foresees the possible alteration of the environment through the introduction by aquaculture of exotic species or through the application of new technologies. States are to be held liable for damage inflicted not only in waters under the jurisdiction of other states but also on the high seas. The appropriate organizations, particularly the Intergovernmental Maritime Organization and the United Nations Environment Programme, are enjoined to get on with the work in which they are already engaged and find their authority thereby reinforced.

Indeed, through its Regional Seas Programme the United Nations Environment Programme has already begun the transformation of the "soft" law of the Convention into "hard" enforceable law. There are now 11 such programs engaging about 110 states and a great number of intergovernmental and nongovernmental organizations [*see illustration on page 49*]. Participating nations are taxing themselves for the costs of cleaning up the seas and preventing further pollution. While the Conference was still in session the United Nations Environment Programme was putting into force the universal standards and norms for pollution control set out in the draft Convention.

Encountering indifference rather than resistance from the North, the developing countries made the Convention articles on marine technology transfer their own. They went as far as they could, but the result is "soft" law indeed. The transfer of technology presents a complex technical, social, political and educational task. One may well ask whether it can be legislated. The law does what it can to set goals and specify modes of cooperation; the rest is up to history. That history is already delineating itself. In the industrial countries the superannuated patent system has small correspondence to the realities of technological competition. Intellectual property is bound to find its natural place in the Common Heritage of Mankind.

The Convention makes an important contribution to the development of a binding dispute-settlement system and thus to the maintenance of peace. It establishes, in fact, the most comprehensive and binding global system ever devised. A new supreme court is established in the International Tribunal of the Law of the Sea that is to have its seat, appropriately, in the Hanseatic city of Hamburg. In one of its special chambers, the Seabed Chamber, corporations and individuals will have standing. Regrettably, but inevitably, there are loopholes in the system. States can claim exemption from compulsory dispute-settlement procedures, particularly with regard to the most sensitive issues, for example the delimitation of sea boundaries or issues arising from the military uses of the sea.

In this, as in so many other respects noted here, the Convention is an ambiguous document. The Law of the Sea does not, however, lead an autonomous existence. It must be seen as the expression of the order and disorder, intranational as well as international, of our time. If present trends continue in other sectors of international relations, states will exploit the loopholes, defects and contradictions of the Convention to plunder and pollute the ocean, exhaust its living resources and deploy their weapons in its depths. If states, on the other hand, turn to serious negotiation of balanced global development and peace, they will find that the Convention provides a foundation for building a better world.



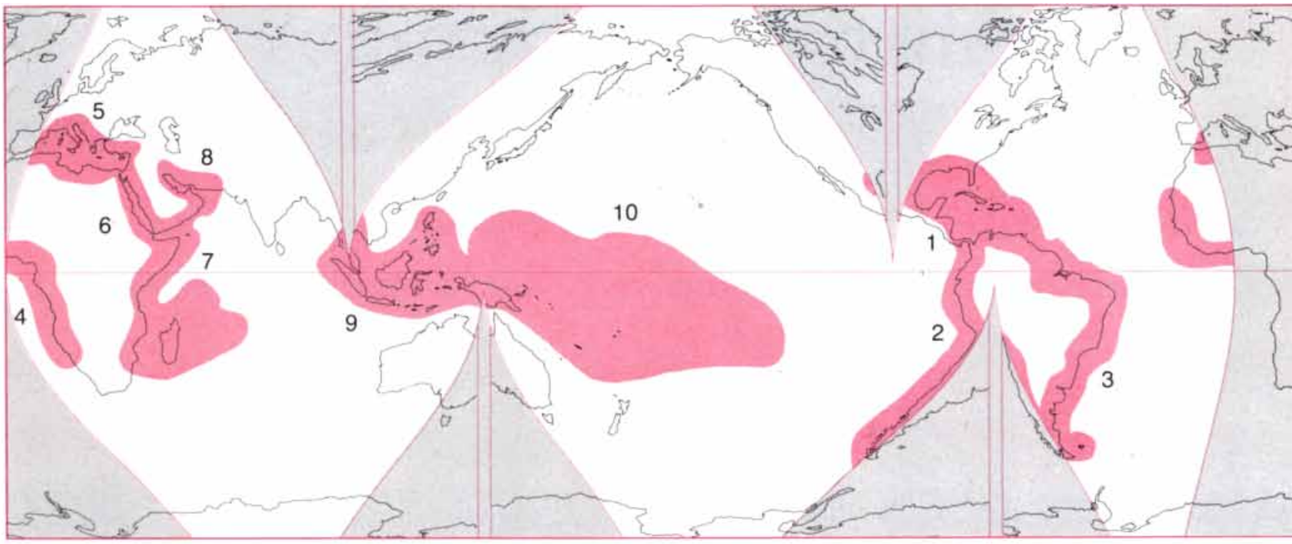
DEEP-SEA MINERAL DEPOSITS in the Gulf of California are revealed by this side-looking sonar image. The features projecting up from the bottom are mounds and "spires" characteristic of hot-spring areas adjacent to sea-floor-spreading centers. In such areas metal-rich sulfides precipitate out of hot water flowing up through the bottom. The features extend 30 meters above the bottom. The distance between four horizontal lines is 450 meters. The image was supplied by Fred N. Spiess and Peter F. Lonsdale of the Scripps Institution of Oceanography.

The ocean is bound to play an ever greater role in the world economy. From small present beginnings in aquaculture the traditional hunting-and-gathering technology will give way to the cultivation of aquatic plants and the husbandry of aquatic animals. The seabed will meet the demand for metals long after the thinning of continental ores has exhausted the ingenuity of technology. By the end of the century the seabed will supply more than half of the world's oil and the ocean will hold energy in reserve in its tides, waves and currents, in its thermal and salinity gradients, in its biomass and prospectively not least in the rare but abundant deuterium atoms of its water molecules.

The ascending economic importance of the ocean will bring it more directly into the political life of nations and of the international community. Nations are creating departments of ocean development, ministries of ocean affairs and the like to assume new functions of ocean management. They are updating their maritime law and extending the rule of law into areas and activities hitherto not covered by law. Harmonizing their practices and institutions with the Convention on the Law of the Sea, they are internalizing and implementing the provisions of the Convention.

The competent international organizations so often invoked in the Convention—the Intergovernmental Maritime Organization, the Food and Agriculture Organization of the United Nations, the Intergovernmental Oceanographic Commission of UNESCO and the United Nations Environment Programme—are taking up their tasks. Particularly for coastal developing countries these agencies are charged with helping to frame national legislation, to improve the management of living resources in their extended economic zones, to chart shipping lanes and schemes for traffic control, to set up regional centers for the advancement of ocean science and technology and to arbitrate conflicts.

Although it seems paradoxical, the expansion of national jurisdiction into the ocean does not diminish the need



"ACTION PLANS" are in effect or in various stages of preparation under the Regional Seas Programme of the United Nations Environment Programme. Plans in effect are those for (*5*) the Mediterranean, (*6*) the Red Sea and the Gulf of Aden and (*8*) the Kuwait Action Plan Region. Plans in preparation are those for (*1*) the Caribbean, (*2*) the Southeast Pacific, (*3*) the Southwest Atlantic, (*4*) West Africa, (*7*) East Africa, (*9*) the East Asian Seas and (*10*) the Southwest Pacific. The projection is the same as that of the map on pages 44 and 45.

for international cooperation; on the contrary, it increases the need greatly. There can be no international cooperation without well-organized nations. In the technological and economic reality of today nations cannot maintain their economic health without strong international cooperation.

The worldwide networks of the Regional Seas Programme are breaking paths to new modes of international cooperation. So important to each nation are its interests in the sea that Turks and Greeks, Israelis and Arabs, Iraqis and Iranians have continued cooperation in their Regional Seas programs even in times of war. Success of a Regional Seas Programme in the Indian Ocean could give impetus to the Sea of Peace concept for that region, which has been adopted and reiterated by the UN but not respected by the superpowers. In the Mediterranean the development of machinery for the funding of the Regional Seas Programme might revive an idea first put forward by Malta and Mexico in the deliberations of the Seabed Committee in 1972: that nations be taxed on their use of the ocean, on revenues from offshore oil, on commercial fisheries and on seaborne trade.

The signing of the Convention at Montego Bay brings into action as of March a new organization, the Preparatory Commission, charged with preparing the rules and regulations for the International Seabed Authority and a provisional agenda for the first session of its governing bodies. The convening of the first session, however, must await the ratification of the Convention. In the interim, the duration of which cannot be predicted, the Preparatory Commission is endowed by the Conference with some of the powers of the Authority itself. Among its considerable powers, functions and responsibilities, the Commission is to examine applications from "pioneer investors" for exploration sites; verify prospecting data and choose the sites to be reserved for the future Authority and its Enterprise, and negotiate arrangements for technology transfer to the Authority and for the training of personnel from developing countries. In sum, the Preparatory Commission has considerable operational capacity. In a way it provides an interim regime for exploration, research and development in ocean mining technology, and it is likely that activities of the Seabed Authority itself will be limited to exploration, research and development until the end of this century.

Once it has acquired the requisite technology—possibly through a joint venture or joint ventures with mining companies—the Commission could apply it not only to explore for nodules in the international area but also to help developing countries that invite such assistance to prospect and develop the resources of their national economic zones. There is nothing in the Convention that would prevent the Authority, or the Commission before it, from performing such needed and useful services, provided they are sought by a coastal state.

Another immediately useful service the Commission could render would be to create marine parks, regions "disapproved for exploitation," and reserve them for scientific purposes. The first such park could well cordon off the Galápagos vents and the unearthly colonies of giant worms and crustaceans sustained there by bacterial chemosynthesis that evolves sulfur rather than oxygen, as in the more familiar and ubiquitous life-sustaining photosynthesis.

A third immediately useful service would relate to arms control and disarmament. The treaty of 1972 banning nuclear weapons from the sea floor made no provision for inspection to verify the observance of its terms. Inspection was left by default to the superpowers. The majority of the countries wanted to internationalize the monitoring and surveillance but were frustrated by the lack of a competent international organization. The Preparatory Commission, and the Authority, could fill the need.

By 1987, when the seabed disarmament treaty next comes up for review, the Commission may well have acquired just the know-how needed; the same acoustic, seismic, magnetic and electronic technologies employed in resource exploration can be turned to the inspection task. An amendment to the treaty in 1987, which would be favored by a large majority of the signatories, could entrust this task to the Preparatory Commission or to the Authority itself if the Convention has been ratified by then.

The signing of the Convention at Montego Bay by the overwhelming majority of the international community thus opens a new phase of the building of an international order. It offers new instruments for enhancing economic development, advancing disarmament, ensuring protection of the environment and creating new forms of scientific and industrial cooperation between North and South. These possibilities would not be there without the new Law of the Sea.

Microprogramming

In most modern computers the routing of information is controlled at the lowest level by a microprogram: a set of stored instructions that functions in place of a completely "hard-wired" control system

by David A. Patterson

A fundamental issue in the design of any computer is how to control, or steer, the electrical signals that represent information. In the arithmetic and logic unit, where the actual processing of information is done, signals must be routed between various counters, adders and other components. The control system must also mediate the transfer of information between the central processor, the main memory units and the various input and output devices. In one approach the control system is completely "hard-wired," that is, it is laid down permanently in the processor's electrical circuitry. A second approach is more flexible and in many cases less expensive. The essential idea is to reduce the complexity of the control system by recording the detailed instructions for controlling the computer in a coded form. In other words, the sequence of paths a signal is to follow is embodied in a program, which is stored in a separate memory unit incorporated into the processor.

In the hierarchy of programs that operate a computer the instructions executed by the control system occupy the lowest and most elementary level; each instruction specifies a single functional state of the machine. Because the control instructions are responsible for such fine details, the task of defining and encoding them is termed microprogramming, thereby distinguishing it from the writing of the higher-level programs known generically as software. A set of control instructions—a microprogram—is said to be written in microcode. The idea of microprogramming was conceived more than 30 years ago, soon after the advent of the first computers. At the time the hardware needed to implement the idea did not yet exist. The method has been adopted, however, in most computers being built today.

In recent years a good deal of confusion has arisen about the meaning of the term microprogramming, owing largely to the advent of the microprocessor, the "computer on a chip" that is at the heart of the latest products of the progressive miniaturization of silicon-based semiconductor technology. It must be emphasized that microprogramming a computer is not the same as programming a microcomputer; in principle any computer, from the largest "mainframe" system to the smallest personal computer, can be designed with a microprogrammed control system. To avoid such confusion microprograms are sometimes classified as firmware, thereby signifying their intermediate status between hardware and software. For clarity I shall adopt the following convention: the term macroinstruction will refer to a software instruction and the term microinstruction will refer to a firmware instruction.

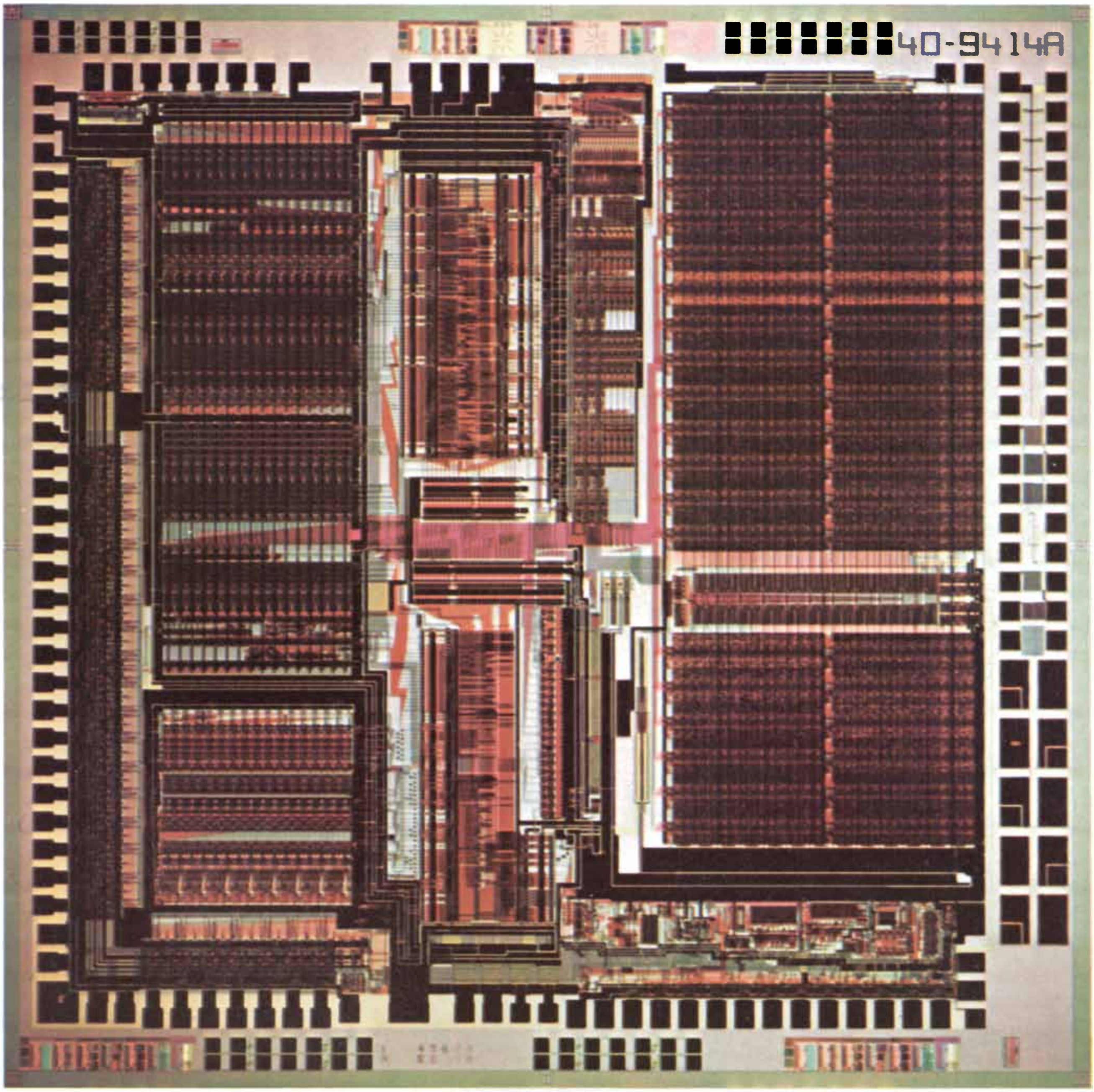
Before microprogramming can be explained it is necessary to review some of the basic principles of computer hardware and computer programming. In many ways a computer is like a simple calculator. Numbers are introduced into both devices by means of a keyboard, and they are read out on a display unit. Some calculators have a memory key to record a number for later retrieval; saving and restoring numbers is also the purpose of the memory unit in a computer. A computer memory, however, is capable of handling millions of numbers. To keep track of so many numbers each one is assigned a specific address.

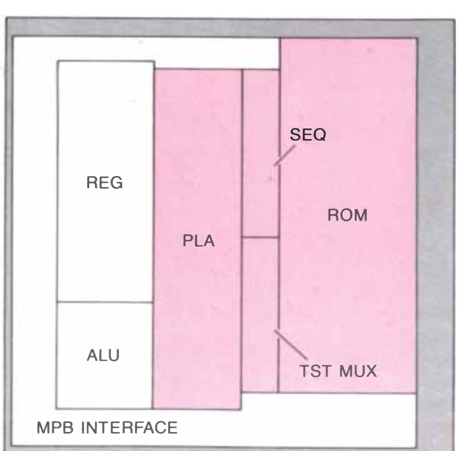
Both a calculator and a computer must be told what to do. Giving a computer a command is analogous to pressing an operator key on a calculator. For example, almost every computer has the macroinstructions ADD, SUB, MULT and DIV, which correspond exactly to the keys marked $+$, $-$, $\times$ and $\div$ on a calculator. Similarly, the macroinstruction LOAD performs the same function as the memory key on a calculator.

Unlike a simple calculator, however, a computer is preeminently a decision-making machine. To understand the distinction consider the sequence of operations represented by the last few sections of Form 1040, the U.S. Individual Income Tax Return [*see top illustration on page 53*]. A computer programmer faced with the chore of completing the form might begin by making a flow chart, which serves as a kind of blueprint for constructing a program [*see illustration at bottom left on page 53*]. A taxpayer working with a simple calculator could follow the flow chart, carrying out the instructions in sequence until he reached the diamond-shaped "decision" step. He would then compare the numbers in the preceding two boxes and, depending on which number was greater, choose the appropriate path to continue his calculations.

For a taxpayer working with a computer rather than a calculator the programmer could go a step further: he could get the computer to make the decision by itself. Following the flow chart, he would first instruct the computer to add successively the numbers entered on lines 50 through 58 of the tax form, storing the sum of each addition at a specific address in the computer's memory; in this case the addresses of the numbers are designated *L*50, *L*51, *L*52 and so on, and they are added successively to an "accumulator" address designated *A* [*see illustration at bottom right on page 53*]. The total for this sequence of steps would then be stored at address *L*59. Next the programmer would do the same for lines 60 through 66, storing the total for the second sequence at address *L*67. At this point (step 19) he would give the computer the following macroinstruction: "JUMPIF $A < L59$, 23." This macroinstruction, called a conditional jump, tells the computer to begin executing a new sequence of macroinstructions at step 23 if the number at the accumulator address is less than the number at address *L*59. If it is not, the program would continue with step 20.

It is the conditional-jump macroinstruction, corresponding to the diamond-shaped box in the flow chart, that endows the computer with the ability to





MICROPROGRAMMED MICROPROCESSOR is at the heart of the Hewlett-Packard Company's HP 9000 family of computers. The microelectronic circuit is the densest of its kind ever made: it has the equivalent of some 450,000 transistors on a square chip of semiconducting silicon 6.3 millimeters on a side. The standard spacing between the signal-carrying lines is one micrometer. The map at the left shows the seven major subsections of the chip. The colored area identifies those subsections that constitute the processor's control system and the white area corresponds to the data path. The control system consists of four parts: a read-only memory (*ROM*), where the microinstructions for controlling the flow of information throughout the computer are stored; a programmed-logic array (*PLA*), where most of the microinstructions are decoded before being transmitted to their various destinations; a sequencing device (*SEQ*), where the order of the microinstructions is arranged, and a test multiplexer (*TST MUX*), where certain conditional microinstructions are decoded. The three other subsections are an array of memory registers (*REG*), where the data being handled by the computer are stored in the form of "words," each of which is 32 bits long; an arithmetic and logic unit (*ALU*), where the actual processing of the 32-bit words is done, and a memory-processor-bus (*MPB*) interface, which oversees communications between the processor chip and the rest of the computer. The rectangular shapes at the top, bottom and right edges of the chip include test devices, special resolution patterns and alignment marks that are unrelated to the operation of the microprocessor.

make decisions, provided of course that the programmer can represent the decisions as simple arithmetic tests. The programmer builds a program by constructing a maze of conditional jumps that leads each input number to its correct destination. On the average one out of 10 macroinstructions in a computer program is a conditional-jump macroinstruction; many computers available today can execute a macroinstruction in less than a millionth of a second, so that a computer typically makes more than 100,000 decisions of this kind per second. The macroinstructions for making the decisions are themselves encoded as numbers and stored along with the other numbers handled by the computer in the main memory. The principle of the stored program, the invention of which was a milestone in the development of the modern digital computer, makes it possible to change the function of a computer by changing the contents of its memory unit instead of by changing its hardware.

As computers get faster they can execute more macroinstructions and make more decisions in a given time span; hence it becomes possible to run increasingly complex programs. Under the circumstances mistakes are difficult to avoid. Programmers prefer a term whose connotations are not quite the same as those of "mistake," however: they say that a program has a "bug." To appreciate the special sense of this usage, consider again the illustration of the income-tax form on the opposite page. Line 68 tells the taxpayer what to do if the amount on line 67 is larger than the amount on line 59, and line 71 explains what to do if line 59 is larger than line 67, but nowhere in the form or in the accompanying instructions is there any indication of what to do if the amounts on lines 59 and 67 are equal! Evidently the Internal Revenue Service has overlooked a bug in its program. It is precisely this kind of detail that a programmer must keep in mind when he is writing and "debugging" a computer program. The difficulty of the task is compounded when it comes to the fine details of writing and debugging a microprogram.

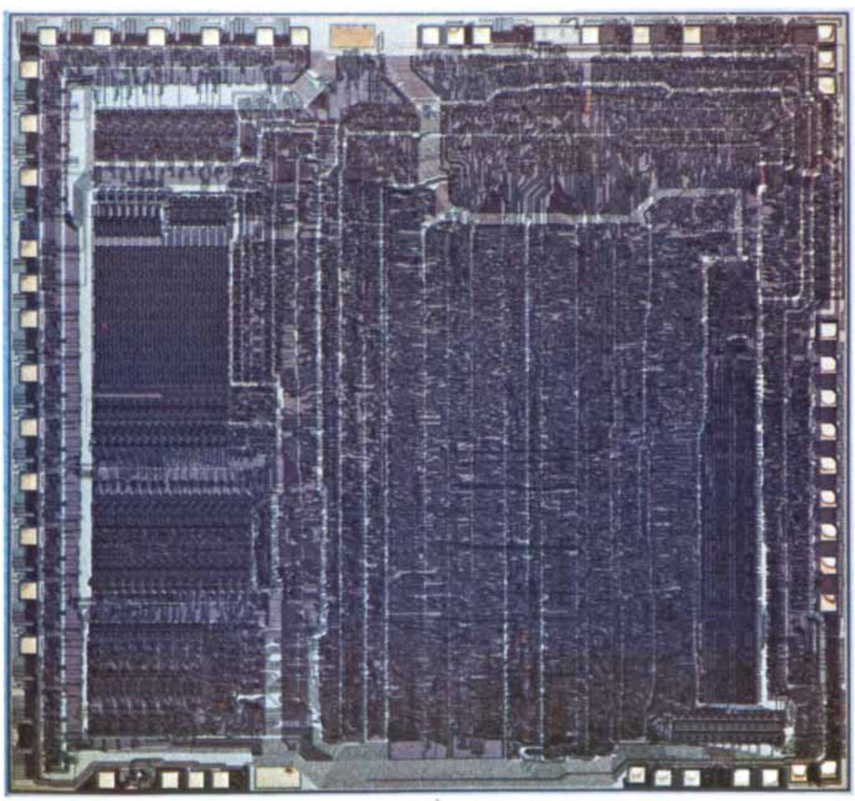
To see how decisions are made and macroinstructions are executed by a computer one must look deeper into the operation of the central processing unit. The processor consists of two major parts: the control system, which makes the decisions and issues the resulting commands to the rest of the hardware, and the arithmetic and logic unit, or data path, where calculations are actually carried out. At the heart of the arithmetic and logic unit is an adder, and in principle all the arithmetic functions can be reduced to repetitive addition steps. Each part of the processor has its own array of short-term memory devices, called registers.

The organization of the computer at this level resembles plumbing. The wires connecting the hardware devices can be thought of as pipes channeling the flow of information. The control system directs the flow by sending an "on" or an "off" signal to open or close each of a large number of valvelike electronic devices called gates [*see top illustration on page 54*]. Because the gates must be opened and closed at the right moment to get the information to the right place, the control signals must be precisely timed. The rhythm of the entire processor is therefore synchronized to an internal clock. Each macroinstruction corresponds to a sequence of clock cycles, with each clock cycle marking a single transfer of information along the data path. The function of the control system is to supply the control signals during the right clock cycles.

In a hard-wired control system a network of electronic logic must be devised to recognize each macroinstruction in the computer's repertory. The macroinstruction itself is a pattern of binary digits, embodied in a pattern of signals sent to the processor. The control system must somehow transform this pattern into the quite different set of signals needed to open and close various gates in the data path. Furthermore, the signals issued by the control system must be timed to coincide with the appropriate clock cycles. Each macroinstruction must give rise to a different sequence of control signals, and so the system becomes more complex as the number of macroinstructions increases. Changes can be made and bugs can be eliminated only by rewiring.

In the early days of the computer the design of the control hardware was a major problem. In comparison the task of designing the other components of the hardware was fairly easy. A teletypewriter could serve as an input-and-output device, and the repetitiveness of the memory structures and of the circuit elements in the data path simplified their design and construction. The hardware needed to count clock cycles and to issue control signals at the appropriate times made the control system quite irregular, and that in turn made the job of designing this part of the central processor much harder.

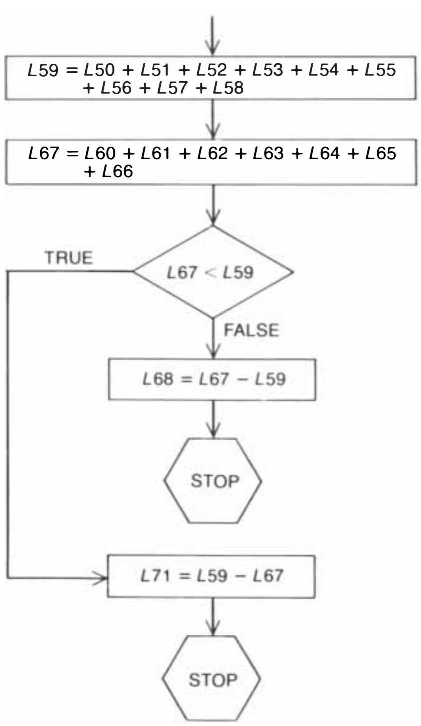
In 1949 Maurice V. Wilkes of the University of Cambridge set out to find a better way to organize the control function. After two years of study he concluded that the best approach is to



NONMICROPROGRAMMED MICROPROCESSOR is distinguished by its highly irregular, hard-wired control section, which occupies the large area at the right in this photograph. The two smaller and more regular areas at the left correspond to the *REG* and *ALU* subsections in the illustration on the preceding page. The irregularity of such a hard-wired control system makes it difficult to design and modify. At the time this chip was made (about five years ago) the designers thought that microprogramming would have required a much larger chip.

	50	**Balance.** Subtract line 49 from line 40 and enter difference (but not less than zero) . ▶	50
Other Taxes (Including Advance EIC Payments)	51	Self-employment tax *(attach Schedule SE)*	51
	52	Minimum tax *(attach Form 4625)*	52
	53	Alternative minimum tax *(attach Form 6251)*	53
	54	Tax from recapture of investment credit *(attach Form 4255)*	54
	55	Social security (FICA) tax on tip income not reported to employer *(attach Form 4137)*	55
	56	Uncollected employee FICA and RRTA tax on tips *(from Form W–2)*	56
	57	Tax on an IRA *(attach Form 5329)*	57
	58	Advance earned income credit (EIC) payments received *(from Form W–2)*	58
06	59	**Total tax.** Add lines 50 through 58 ■	59
Payments Attach Forms W–2, W–2G, and W–2P to front.	60	Total Federal income tax withheld 60	
	61	1982 estimated tax payments and amount applied from 1981 return . 61	
	62	Earned income credit. If line 33 is under $10,000, see page 15 of Instructions 62	
	63	Amount paid with Form 4868 63	
	64	Excess FICA and RRTA tax withheld (two or more employers) . 64	
	65	Credit for Federal tax on special fuels and oils *(attach Form 4136)* 65	
	66	Regulated Investment Company credit *(attach Form 2439)* 66	
	67	**Total.** Add lines 60 through 66 . ▶	67
Refund or Amount You Owe	68	If line 67 is larger than line 59, enter amount **OVERPAID** . ▶	68
	69	Amount of line 68 to be **REFUNDED TO YOU** . ▶	69
	70	Amount of line 68 to be applied to your 1983 estimated tax . . . ▶ 70	
	71	If line 59 is larger than line 67, enter **AMOUNT YOU OWE.** Attach check or money order for full amount payable to Internal Revenue Service. Write your social security number and "1982 Form 1040" on it. ▶ (Check ▶ ☐ if Form 2210 (2210F) is attached. See page 16 of Instructions.) ▶ $	71

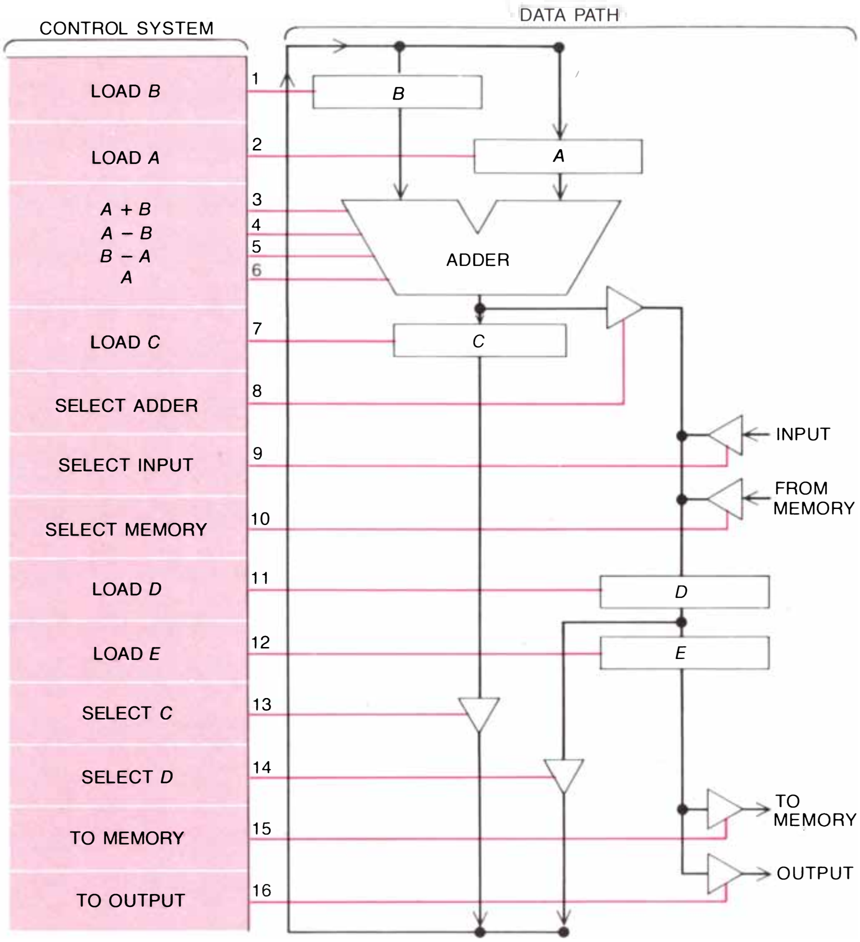
FAMILIAR FORM serves as an example of a sequence of operations that can be done by a computer program. The illustration reproduces the last 22 lines of Form 1040, the U.S. Individual Income Tax Return for 1982, prepared by the Internal Revenue Service.



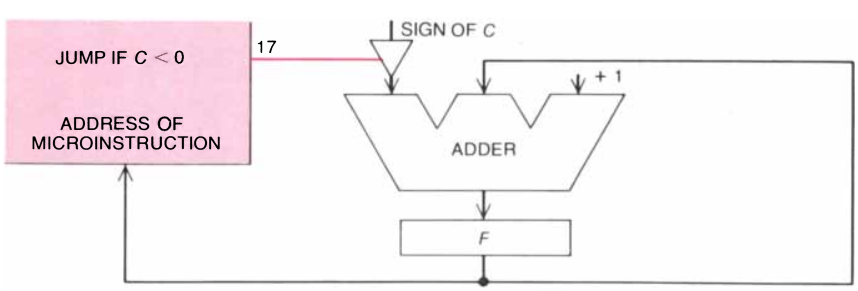
ADDRESSES	INSTRUCTIONS	COMMENTS
1	LOAD *L*50, *A*	*A* = *L*50
2	ADD *L*51, *A*	*A* = *L*50 + *L*51
3	ADD *L*52, *A*	*A* = *L*50 + *L*51 + *L*52
4	ADD *L*53, *A*	*A* = *L*50 + *L*51 + *L*52 + *L*53
5	ADD *L*54, *A*	*A* = *L*50 + *L*51 + *L*52 + *L*53 + *L*54
6	ADD *L*55, *A*	*A* = *L*50 + *L*51 + *L*52 + *L*53 + *L*54 + *L*55
7	ADD *L*56, *A*	*A* = *L*50 + *L*51 + *L*52 + *L*53 + *L*54 + *L*55 + *L*56
8	ADD *L*57, *A*	*A* = *L*50 + *L*51 + *L*52 + *L*53 + *L*54 + *L*55 + *L*56 + *L*57
9	ADD *L*58, *A*	*A* = *L*50 + *L*51 + *L*52 + *L*53 + *L*54 + *L*55 + *L*56 + *L*57 + *L*58
10	STORE *A*, *L*59	*L*59 = *L*50 + *L*51 + *L*52 + *L*53 + *L*54 + *L*55 + *L*56 + *L*57 + *L*58
11	LOAD *L*60, *A*	*A* = *L*60
12	ADD *L*61, *A*	*A* = *L*60 + *L*61
13	ADD *L*62, *A*	*A* = *L*60 + *L*61 + *L*62
14	ADD *L*63, *A*	*A* = *L*60 + *L*61 + *L*62 + *L*63
15	ADD *L*64, *A*	*A* = *L*60 + *L*61 + *L*62 + *L*63 + *L*64
16	ADD *L*65, *A*	*A* = *L*60 + *L*61 + *L*62 + *L*63 + *L*64 + *L*65
17	ADD *L*66, *A*	*A* = *L*60 + *L*61 + *L*62 + *L*63 + *L*64 + *L*65 + *L*66
18	STORE *A*, *L*67	*L*67 = *L*60 + *L*61 + *L*62 + *L*63 + *L*64 + *L*65 + *L*66
19	JUMP IF *A* < *L*59, 23	GO TO ADDRESS 23 IF *A* < *L*59 (SAME AS *L*67 < *L*59)
20	SUB *A*, *L*59	(*L*67 ≥ *L*59) THEREFORE *L*67 − *L*59
21	STORE *A*, *L*68	*L*68 = *L*67 − *L*59
22	STOP	FINISHED
23	LOAD *L*59, *A*	(*L*67 < *L*59) THEREFORE *L*59 − *L*67
24	SUB *L*67, *A*	*A* = *L*59 − *L*67
25	STORE *A*, *L*71	*L*71 = *A* (= *L*59 − *L*67)
26	STOP	FINISHED

FLOW CHART serves as a kind of blueprint of the sequence of operations needed to complete the last 22 lines of the income-tax form. The amount entered on each line of the form is stored at a specific address in the computer's memory; in this case the addresses of the numbers are designated *L*50, *L*51, *L*52 and so on. The diamond-shaped box is computer programmer's symbol for a "decision" step.

PROGRAM OF MACROINSTRUCTIONS for completing the income-tax form follows the flow chart. Each macroinstruction is itself stored at a memory address (*first column at left*). The numbers at addresses *L*50, *L*51, *L*52 and so on are added successively to an "accumulator" address designated *A*. Steps 1 through 10 correspond to the topmost box in the flow chart and and steps 11 through 18 correspond to the second box from the top. The conditional-jump macroinstruction at step 19 ("JUMPIF *A* < *L*59, 23") is equivalent to the diamond-shaped decision box. Steps 20 and 21 correspond to the box on the "False" path of the flow chart and steps 23, 24 and 25 correspond to the box on the "True" path. The comments at right are notes to remind the programmer of what is going on; they do not affect the execution of the program.



IDEALIZED CENTRAL PROCESSOR relies on 16 control wires (*colored lines*) to guide the electrical signals representing information through a simplified data path (*black lines*). In this case the arithmetic and logic unit has only one special-purpose device: an adder, which receives numbers from two input wires, adds or subtracts them in accordance with commands received over control wires *3* through *6* and sends the result through an output wire. The rectangular shapes are registers, short-term memory components that can hold one word of information at a time. The triangular shapes are gates, devices that pass the information along when the associated control wire is on and block the information when the control wire is off. In a nonmicroprogrammed processor each of the commands listed in the control section is defined by specific control hardware. In the microprogrammed version of the same processor each command is stored as a separate microinstruction in the microprogram memory. A sample microprogram for this simplified processor is given in the top illustration on the opposite page.



ADDITIONAL HARDWARE IS NEEDED to enable the microprogrammed version of the simplified processor depicted at the top of the page to make decisions of the conditional-jump type. Register F holds the address in microprogram memory of the microinstruction currently in control of the data path. The function of the extra adder is to add one ("+1") to the address of the current microinstruction, thereby transmitting the address of the next sequential microinstruction in the microprogram. The adder also feeds information from the data path back into the control system. The sign of the number in register C goes through a gate controlled by wire *17* into the third input of the adder. When this wire is off, the microinstruction at $F + 1$, the next sequential address, is executed (since $0 + F + 1$ is equivalent to $F + 1$). The adder adds either $0 + F + 1$ or $1 + F + 1$ depending on whether number in register C is respectively positive or negative. Thus with the aid of control wire *17* the microprogrammer can conditionally jump over a microinstruction, depending on the sign of the number in C. The microprogram for the conditional-jump operation is given in the bottom illustration on the opposite page.

think of the control system as a matrix, or rectangular grid, with each row of squares corresponding to a clock cycle and each column associated with a control wire. For example, the simplified processor shown in the top illustration at the left has 16 control wires, and so the control matrix corresponding to it is 16 squares wide. Choosing a sequence of operations then becomes a simple matter of putting a suitable combination of binary symbols in the squares: for each control wire that must be on during a particular clock cycle write a 1 in the corresponding square, and for each control wire that must be off write a 0.

The hardware equivalent of the control matrix is a simple, repetitive memory structure. The content of each cell in a row determines the state of the corresponding control line for the duration of one clock cycle. The pattern of digits that constitutes a macroinstruction serves merely to select the appropriate row (or sequence of rows) in the control memory. In other words, the macroinstruction becomes an address designating the row. Designing a control system then becomes more a matter of software than of hardware; the complexity lies in choosing the right pattern of 1's and 0's to store in the memory rather than in specifying the right combination of hard-wired circuits to generate the control signals.

Wilkes recognized the similarities between his approach and conventional programming, and so he borrowed terms from programming to describe his ideas, in each case adding the prefix "micro-" to indicate the elementary nature of the control function. Hence "microprogramming" and its family of related terms. In particular Wilkes named each row of squares in the control matrix a microinstruction, and each sequence of rows that executes a single macroinstruction he called a microprogram. The memory structure designed to hold the microprogram is called the microprogram memory. The concept of microprogramming made it easier to understand the control function and, by replacing the complex control circuitry with a repetitive array of memory cells, made it easier to build the hardware. Most important, it gave the computer a new flexibility: the control system could now be changed without redesigning the hardware.

Consider a sample microprogram for the simplified processor discussed above [*see top illustration on opposite page*]. Suppose the operation controlled by the microprogram is to subtract the number stored in memory register B from the one in the accumulator register A and then to store the result in A. By examining the data path in the hardware diagram one can see that the device labeled "adder" is capable not only of

CONTROL WIRES

MICROINSTRUCTIONS	1 LOAD B	2 LOAD A	3 A + B	4 A − B	5 B − A	6 A	7 LOAD C	8 SELECT ADDER	9 SELECT INPUT	10 SELECT MEMORY	11 LOAD D	12 LOAD E	13 SELECT C	14 SELECT D	15 TO MEMORY	16 TO OUTPUT	COMMENTS
1	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	LOAD *A* − *B* IN *C;* GO TO 2
2	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	LOAD *C* IN *A*
3																	CONTINUE

SAMPLE MICROPROGRAM for a single macroinstruction that can be executed by the simplified processor shown in the top illustration on the opposite page consists of a sequence of rows in a control matrix, or rectangular grid, in which each row corresponds to a microinstruction and each column is associated with a control wire. For each control wire that must be on during a particular clock cycle a 1 is written in the corresponding square and for each control wire that must be off a 0 is written. (Here "on" is also signified by color and "off" by white.) The operation controlled by the microprogram is to subtract the number stored in memory register *B* from the one in the accumulator register *A* and then store the result in *A*. As in the case of macroinstructions, the comments at the right are just reminders.

CONTROL WIRES

MICROINSTRUCTIONS	1 LOAD B	2 LOAD A	3 A + B	4 A − B	5 B − A	6 A	7 LOAD C	8 SELECT MEMORY	9 SELECT ADDER	10 SELECT INPUT	11 LOAD D	12 LOAD E	13 SELECT C	14 SELECT D	15 TO MEMORY	16 TO OUTPUT	17 JUMP IF C < 0	COMMENTS
1	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	LOAD *A* − *B* IN *C;* GO TO 2
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	IF *C* < 0 GO TO 4
3	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	LOAD *D* IN *E;* GO TO 4
4																		CONTINUE

DECISION-MAKING MICROPROGRAM utilizes the augmented control hardware shown in the bottom illustration on the opposite page. The operation controlled by the microprogram is to load the number in register *D* into register *E* if *A* is greater than or equal to *B*. The conditional jump calls for an additional square to give the command to jump to microinstruction *4* if the number in register *C* is negative (column 17). Thus the conditional-jump macroinstruction comes down to a choice between microinstruction *3* and microinstruction *4*.

adding the two numbers but also of subtracting either one from the other. Before the result can be put back in *A,* however, it must be held temporarily in register *C;* hence register *C* must be loaded and then unloaded into the data path that leads back to *A*. The subtraction operation calls for two microinstructions. The first microinstruction tells the adder to subtract the number in *B* from the one in *A* and to put the result into *C;* accordingly control wires 4 ("*A* − *B*") and 7 ("LOAD *C*") are turned on. The next microinstruction moves the information from *C* into *A* by opening a gate in the wire from *C* in order to pass the information along to *A;* hence control wires 13 ("SELECT *C*") and 2 ("LOAD *A*") are turned on.

Now consider a somewhat more complicated microprogram for the same processor. The second microprogram calls for a decision-making process similar to the conditional jump in the income-tax program. The decision comes down to a choice between two microinstructions. For this purpose one more square must be added to the microinstructions and one more control wire must be added to the processor. The augmented control hardware is shown in the bottom illustration on the opposite page and the revised microprogram in the bottom illustration on this page.

The operation controlled by the second microprogram is to load the number in register *D* into register *E* if *A* is greater than or equal to *B*. First *B* is subtracted from *A* and the result is stored in *C*. If the number in *C* is negative, it follows that *A* is less than *B* and so the decision is to jump over the next microinstruction and resume the microprogram at microinstruction 4. If the result of the subtraction is a positive number or zero, on the other hand, *A* is greater than or equal to *B,* and the microprogram proceeds to the next sequential microinstruction, which loads *E* from *D*. A square is added to the microinstructions to give the command to jump to microinstruction 4 if *C* is negative, and a corresponding control wire is added to the processor's hardware.

A macroinstruction is said to be "interpreted" by the microcode to yield the control signals that actually manipulate the flow of information in the data path. Thus the sequence of microinstructions in the second example above is the microcode interpretation of the conditional-jump macroinstruction written as "JUMPIF" at step 19 of the income-tax program. The microprogram deals with the simple arithmetic test presented by the programmer by directing the hardware to test the number in the *C* register to see if it is negative, and if it is, to execute the appropriate conditional-jump microinstruction. For the hardware to perform such a test it must feed information from the data path back into the control system. The hardware needed to determine whether the number in *C* is negative is simply a wire from *C* that represents the sign of the number in that register: the wire is off (0) when the sign is positive (implying that *C* is greater than or equal to zero), and it is on (1) when the sign is negative (implying that *C* is less than zero).

The substitution of simple, repetitive memory structures for complex hard-wired control circuits yields two main advantages: it makes it easier to understand and build the control system, and it makes it easier to modify the system. One can correct a mistake in a microprogrammed control system simply by changing the contents of the memory. For example, suppose one decides to change the preceding microprogram from one that loads E from D if A is less than B to one that loads E from D if B is less than A. The first microinstruction in the original microprogram loads C with the result of $A - B$. If one turns off control wire 4 and turns on control wire 5, however, then the microinstruction would load C with the result of $B - A$. In either case the next microinstruction determines whether the number in C is less than zero and jumps if it is. Hence one can correct the operation by changing only two control wires in one microinstruction. With a nonmicroprogrammed control system one would have to change the structure of the hardware to make the same correction.

Besides reducing the cost of correcting the control system the introduction of microprogramming makes it possible to change the control system completely by changing the total contents of the microprogram memory. Thus a single hardware system can be made to serve many different functions.

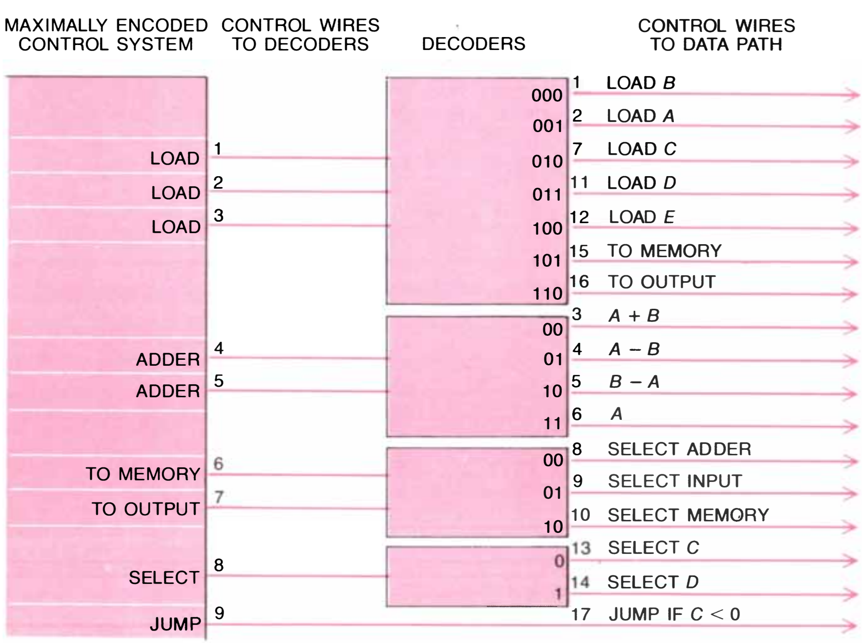
The advantages of microcode were apparent almost from the beginning. Wilkes's idea, however, was somewhat ahead of the hardware technology. Microprogramming calls for fast and inexpensive memory structures, and they were not available in the 1950's. Furthermore, it was generally thought at the time that any computer designed to take advantage of this ingenious approach would be slower than a machine with a hard-wired control system. As a result microprogramming remained little more than an intellectual curiosity for more than a decade.

It was not until the early 1960's that microprogramming emerged into the limelight, largely as the outcome of a major business decision by the International Business Machines Corporation. In the late 1950's IBM had in effect become many small companies, each one marketing its own type of computer. The computers differed not only in size, speed and price but also in the set of macroinstructions provided to operate them. Because of the differences in macroinstruction sets, programs devised for one type of IBM computer could not be run on another IBM computer. Obviously it would have been much easier to sell computers that operated with standardized software and that differed only in size, speed and price. The problem was to convince the managers of the subsidiary "companies" that it was possible for both the designers of the fastest computers and the designers of the cheapest computers to build efficient machines that relied on the same set of macroinstructions.

As it happened, the managers of the subsidiaries agreed to the innovative concept of a family of software-compatible computers only because they were persuaded that microprogramming made such a plan feasible. If one could microprogram the same hardware to execute different sets of macroinstructions, it was reasoned, then one should be able to microprogram different computers to execute the same macroinstruction set. The perceived marketing advantages of the family concept were so great that IBM made a major investment in the development of memory technologies that could lead to fast and inexpensive microprogram memories. Finally, in 1964 IBM announced a family of seven software-compatible computer models that varied in speed by a factor of 300 and in cost by a factor of 100.

Almost as an afterthought it was decided at IBM that microprogramming could help with the difficult task of making the programs for the old computers compatible with the new ones. Accordingly several models in the new IBM family ended up with two microprograms: one for the new, improved set of macroinstructions and another for the old macroinstruction set. This approach enabled customers to run their old programs without modification on the new hardware. The method proved so popular that for the first few years it was difficult to say whether the new IBM machines were spending more time running the microprogram for the new macroinstructions or the microprogram that emulated the old ones.

Most major computer manufacturers soon followed the lead of IBM in relying on microcode to produce a family of software-compatible machines that vary over a wide range of speeds and prices. The main difficulty encountered along the way was in duplicating the expensive memory technology invented by the engineers at IBM. Rapid progress in microelectronic technology eventually solved the problem. Every year for more than a decade the semiconductor industry was able to double the amount of information that could be stored on a single memory chip. This technology solidified the position of microprogramming as the control method of choice through the 1970's. The exceptions were "supercomputers," the fastest and most expensive processors. In that case special-purpose control is still



NUMBER OF CONTROL WIRES can be reduced by further encoding the control signals and introducing suitable decoding devices. In general this approach involves finding two wires that are rarely both on for the same microinstruction and eliminating one of them, making the other serve a dual function. This schematic hardware diagram depicts a maximally encoded version of the processor shown in the two illustrations on page 54. The control wires to the data path have been rearranged to emphasize the wires that have been encoded together in the new control system. The additional encoding generally results in a control system that is slower than a minimally encoded one (in other words, one in which each microprogram square is matched by one control wire), because fewer operations can proceed simultaneously and because the control signals must pass through the decoders. In addition the resulting microprogram tends to have a larger number of microinstructions. Because each microinstruction has fewer binary digits, however, the cost of microprogram memory can be significantly reduced.

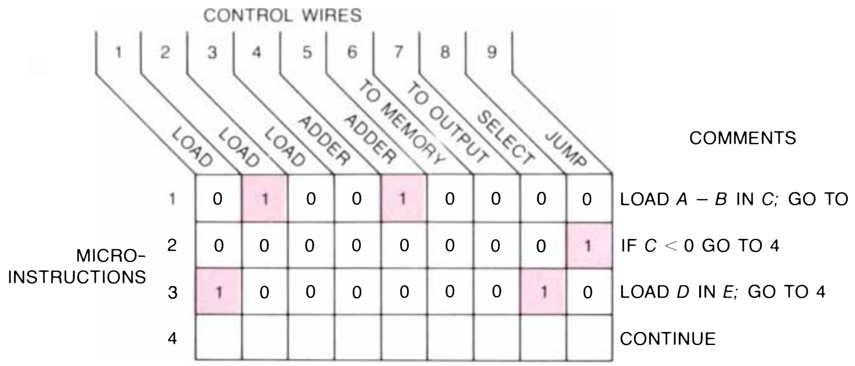
preferred for its greater speed, even though it is harder to build.

When microprocessors began to appear on the scene a decade or so ago, the question naturally arose of how to design their control systems. Repeating history, the first microprocessors were built with nonmicroprogrammed control systems. As the technology has matured, however, the advantages of microprogramming have persuaded most microcomputer designers to choose microcode.

In the meantime one kind of microprogramming has evolved away from the simple matrix concept in which each square is matched to an individual control wire. The improvements have come mainly from the effort to reduce the cost of the microprogram memory. In the execution of any single microinstruction it is generally the case that only a few components of the hardware are active; the control wires leading to all the other elements are off. As a result a microprogram tends to have islands of 1's scattered in a sea of 0's. If the same end could be achieved with less memory (that is, with fewer 0's), the result would be a less expensive memory. The solution is to supply the same information to the data path by further encoding the control signals. In general the additional encoding step involves finding two wires that are rarely both on for the same microinstruction and eliminating one of them, making the other serve a dual function. This approach tends to result in a larger number of microinstructions, but each microinstruction has fewer binary digits. A maximally encoded set of microinstructions is said to be "vertical," because the resulting microprograms are usually tall and narrow. A minimally encoded set of microinstructions is said to be "horizontal," because the resulting microprograms are usually short and wide. Minimally encoded microinstructions generally lead to faster computers, since more operations can be done simultaneously.

The growing importance of microprogramming has also stimulated interest in finding ways to enable computer users to write their own microcode. The potential advantages of transforming the computer from a general-purpose problem solver to one specifically tailored to a given application has made user microprogramming an attractive option. So far the best results obtained by users who have written their programs in microcode have shown an improvement in speed by as much as a factor of 10. In practice the gain in speed depends on the particular data path of the microprogrammed computer and on the application; in some cases the speed advantage is much less, and occasionally it may even disappear.

In most microprogrammed computers the microinstructions are stored in a form of read-only memory, whose contents are determined when the memory array is manufactured. Another kind of microprogrammed computer is capable of changing the contents of its microprogram memory in as little as a few thousandths of a second, while the computer is in operation. Such a machine is said to have a writable control store. The key to this approach is to have many different microcoded applications available to be loaded only when necessary. Endowing a computer with a writable control store is a way of removing the hardware barrier from user microprogramming.



MAXIMALLY ENCODED MICROPROGRAM for the control system on the opposite page is said to be "vertical," because it usually has a larger number of shorter microinstructions, thereby making it taller and narrower than the corresponding "horizontal" microprogram. In this case the encoding was done in such a way as to require no additional microinstructions.

In spite of these efforts microprogramming has remained so difficult that few users have taken full advantage of its potential. Therefore some investigators have attacked the problem of making it easier to write microcode. In many ways the situation is similar to one that arose earlier in the history of computer programming. Then the problem was the difficulty of arranging thousands of macroinstructions into a program that would operate without fail. This problem inspired programmers to write special programs to help write the application programs. The special programs translate high-level computer languages into the simpler macroinstructions a computer can execute. The high-level languages such as BASIC and Pascal are easier to understand and use, thereby making the task of programming less difficult, albeit at some sacrifice in the speed of the program.

For the past few years workers in microprogramming have been trying to create analogous translation programs for microcode. The problem is harder in this case because there are many more potential microinstructions than there are ordinary macroinstructions. A microprogrammed processor might have 50 control wires; since each wire can have one of two states (on or off), the total number of possible microinstructions would be 2^{50}, or approximately 10^{15}. Moreover, since the speed of the microprogram determines the speed of the computer, the demand for efficiency is much higher. The result has been the invention of high-level microprogramming languages aimed at balancing ease of microprogramming with efficiency. Recent developments have been encouraging, but general acceptance among the practitoners of microprogramming is still several years away.

Meanwhile investigators at the IBM Thomas J. Watson Research Center and my colleagues and I in the department of electrical engineering and computer science of the University of California at Berkeley have taken a different approach. We start by discarding the conventional macroinstruction set. We then write new programs that translate the standard high-level programming languages directly into microcode. Machines based on this principle are known generically as reduced-instruction-set computers, or RISC's. To test our ideas our group at Berkeley has designed our own microprocessor, which we call RISC I. In addition we have simplified the translation step by maximally encoding the microinstructions to achieve a highly vertical type of microprogram. In effect, all programs then become microprograms.

Recent tests by our group and by the workers at IBM have shown that a reduced-instruction-set computer would be from three to five times as fast as a conventional complex-instruction-set computer for many applications. Workers at Stanford University and at several industrial laboratories have begun projects to investigate this approach. If the preliminary findings are substantiated, a simpler type of computer might emerge, representing an unexpected advance in the continuing exploitation of the principle of microprogramming.

The Script of the Indus Valley Civilization

One of the four earliest civilizations has a script that has long resisted decipherment. A start on deciphering it is now made on the basis of facts and inferences from the archaeological record

by Walter A. Fairservis, Jr.

The oldest civilizations are those of Sumer, Egypt, China and the Indus Valley in the northwestern part of the subcontinent now occupied by India, Pakistan and Bangladesh. The writings of the first three civilizations can be read; the inscriptions of the fourth one remain largely enigmatic. As a result what is known of the Indus Valley civilization comes solely from the material objects brought to light by archaeology. It is ironic that in a part of the world noted for the antiquity of its literature even the most ancient accounts contain no valid reference to the first great culture that thrived there.

The culture of the Indus Valley civilization is called Harappan, after one of its two great cities: Mohenjo-Daro and Harappa. The difficulties that face those who would decipher Harappan writing seem at first virtually insurmountable. They derive in large part from the Harappans' limited use of their script. Their "texts" consist almost exclusively of brief inscriptions on seals and equally limited graffiti on pottery. No known inscription consists of more than 21 signs and the average text numbers only five or six. This, coupled with the fact that many of the signs are pictographic, has let the imagination of more than a few scholars run riot. Earnest attempts have been made to relate Harappan to Minoan, Canaanite, Hittite and even to the peculiar "writing" of Easter Island. Studies in recent years building on what the archaeological record reveals about Harappan life have now yielded a more rewarding approach to the decipherment problem. It is my intent to outline some of these achievements here.

In the decades since the two great Harappan cities were discovered in the 1920's, and particularly since the end of World War II, archaeologists in India and Pakistan have located nearly 1,000 other Harappan sites. They are spread in a wide arc from western India in the vicinity of the Narmada River northward across Gujarat and Kutch, through the Pakistani regions of Sind and the western Punjab and on into Indian Rajasthan and the eastern Punjab up to the vicinity of Delhi. Other settlements have been found along the coast of the Arabian Sea almost as far as the Iranian border and in Baluchistan; one settlement was even discovered close to the Oxus River, deep in central Asia, by a recent French archaeological mission.

Most Harappan sites are small, covering between two and five acres, and are near rivers or streams. There appear to have been three phases of settlement. Sites of the early phase are in or near the borderlands between the subcontinent and Iran. Those of the mature phase are more widespread. Most are in the valley of the Indus River itself, but such far-flung sites as those in Baluchistan and the one near the Oxus also belong to the mature phase. Those of the later phase tend to be far to the south and east of the Indus. Findings at these later sites also contain evidence that the older Harappan culture was in the process of merging with local pastoral and agricultural peoples. Such findings suggest Harappan civilization did not come to a sudden end but played a role in the development of the style of village life that is so characteristic of the subcontinent today.

What does the archaeological record tell us about Harappan life? First, it indicates that its mainstay was farming. Both grains and garden vegetables were grown, and cattle husbandry was central to the farm economy. Second, it indicates that the material culture of the Harappans was simple but not lacking in rich goods. They used copper and bronze for some tools and weapons and occasionally worked gold and silver into the beads that were their principal form of jewelry. At the same time the bulk of their artifacts consist of wood and bone, shell, flint and clay, all materials that were locally abundant. Among the beads, however, were some skillfully fashioned out of rarer semiprecious stones such as agate, carnelian and lapis. (The site near the Oxus was also close to early lapis mines.)

A further striking aspect of Harappan life was the extent of standardization. In architecture, building bricks were standard in size and were laid in standard ways. Drainage and sewer systems were standard in pattern. Dwellings were standard in dimensions and special structures (possibly public) were positioned with respect to private ones according to standard plans. Other aspects of the same phenomenon included standard weights and measures, pottery that was standard in shape and ornamentation and standardized artifacts such as ladles, loom weights and even toy carts. At the same time certain aspects of Harappan life suggest the later culture of India, for example the use of distinctive headdresses and of multiple bangles and necklaces and even the style of Harappan figurines.

So much for a broad summary of what is known. What is not known may be of equal significance. There is no evidence of rivalry between different Harappan states, of warfare, of major international trade or of the kings and courts and great temple complexes so characteristic of the other ancient civilizations of the Old World. The archaeological evidence reveals next to nothing about Harappan religion or political and social organization. Yet Harappan civilization was important to the civilizations that succeeded it. The Harappans cultivated cotton and perhaps rice, domesticated the chicken and may have invented the game of chess and one of the two great early sources of nonmuscle power: the windmill. (The other was the water wheel.)

When did this enigmatic civilization flourish? Although there is some controversy on the point, the mature Harappan phase seems to have extended from about 2200 to 1700 B.C. Most Harappan sites appear to have been occupied for no more than 200 years; they give the impression of a short-lived development characterized by substantial group organization and regular interaction between contemporaneous settlements.

The task of "reading" the Harappan inscriptions is sufficiently difficult to make many scholars believe it is impossible. The first difficulty arises from the fact that the inscriptions represent the unknown writing of an unknown language for which there are no bilingual texts, such as the Rosetta stone of Egypt or the Behistun monument of Iran. A second difficulty is the absence of long texts. The entire corpus of Harappan writing consists of some 4,000 seals, seal impressions (that is, impressions of seals) and graffiti on pottery; moreover, many of the seals are damaged and many of the pottery graffiti are interrupted by breakage. Still a third difficulty arises from the fact that Harappan civilization was not only geographically remote from the other civilizations of its time but also historically remote from later cultural developments on the subcontinent.

Fortunately the Harappan seals are inscribed with both writing and pictures, and the pictorial motifs can yield clues to what the writing means. Most Harappan seals are square or rectangular pieces of soapstone, a material that is easy to carve. On the back of most of the seals is a raised boss pierced with a hole for a carrying string; on the front is a combination of a negative picture and an inscription. The carvings presumably

SEAL OF THE HARAPPAN CULTURE in the archaeological collection of the National Museum of Pakistan in Karachi displays a zebu bull. The four signs across the top of the seal are among those that appear on the grid on pages 64 and 65. They are, from left to right, the triple stroke of the numeral 3, which also has the meaning "foremost"; a pipal-leaf figure and an arrow, with the meaning "mother" or "mistress"; a stick figure with "horns," indicative of a high-ranking personage, and a pot with handles, an honorific suffix with the meaning "high" or "superior." The seal evidently belonged to a high-ranking member of Harappan society, possibly a woman.

"LORD OF THE BEASTS," a famous seal in the archaeological collection of the National Museum of India in New Delhi, shows a human figure seated cross-legged and wearing a water-buffalo headdress. The "beasts" associated with the figure are a rhinoceros and a water buffalo (*left*), an elephant and a tiger (*right*) and a goat with its head turned (*bottom*). The goat may originally have been one of two, the second being lost when the seal was broken. The inscription can be rendered as "The Black One, the Black Buffalo *an-i(l)*, the High One, the Lord of Chiefs."

GRAFFITI ON POTTERY, such as those on these sherds found in Baluchistan, are a second source of Harappan signs. They are not as numerous as the inscriptions on seals and seal impressions, but they indicate that the inscriptions were meant to be read from right to left.

identified the owner of the seal, so that when the seal was pressed into soft clay, the imprinted object was recognized as somehow associated with the owner.

A majority of Harappan seals display one or the other of two distinct pictorial motifs. The first involves an animal. It is usually a long-horned bull, but humped zebus are also portrayed, as are water buffaloes, goats, short-horned bulls, rhinoceroses, tigers, gavials (the river crocodiles of India) and elephants. Whatever the animal, an object was shown in place before it. The object shown with wild or dangerous animals is platterlike. The object shown with the domestic animals is either a basket or (particularly in front of the long-horned bull) a "stem" emblem such as one also shown being carried in processions.

A few of the animal seals show their subjects in groups. One well-known example is centered on an anthropomorphic figure, sometimes called the "Lord of the Beasts," seated cross-legged and wearing a water-buffalo headdress. To the right of the figure are an elephant and a tiger; to the left are a rhinoceros and a water buffalo. The seal is damaged but below and to the left of the central figure's feet can be seen a goat looking over its shoulder and to the right are what may be the horns of a matching goat figure, otherwise obliterated. A less elaborate seal centers on a pipal, or sacred fig tree (*Ficus religiosa*); the heads of two long-horned bulls are shown growing up from its trunk.

Another "animal group" depiction appears on one prismatic seal. (A few Harappan seals were made in the shape of a prism and a few others were cylindrical.) In a row on one of its three faces are an elephant, a rhinoceros and a tiger; a fourth figure in the row is too worn to identify. Presumably this is the "wild animal" face of the seal. On the second face is another four-animal procession, none of them apparently "wild." On both faces, however, a fifth animal is depicted above the animal procession. It is a gavial, self-evidently a member of the "wild" group. On the "wild animal" face the gavial is shown with a recognizable fish in front of its snout and what may be a fish behind it. On the second face no recognizable fish appears.

A fourth seal, even more elaborate than the "Lord of the Beasts," introduces a "worshiper" element. It repeats the depiction of a pipal tree, this time at its upper right corner. Between the branches of the tree stands a horned anthropomorphic figure. Facing the horned figure is a kneeling one, skirted and thus presumably female; to the left of the kneeling figure is a large goat. Seven skirted figures occupy the bottom half of the seal, their hair dressed in some kind of long "ponytail." A rather gruesome depiction on a fifth seal shows several similarly coiffed figures, one of them wearing a skirt, being attacked by a water buffalo.

These motifs suggest something about Harappan social organization: that individual seal-bearers belonged to groups that transcended normal familial lines. For example, all individuals with a rhinoceros on their seal may have had some social tie in common. What was held in common could have been membership in some superfamilial group such as a clan or a club. The existence of procession scenes on a few seals, where animal effigies seem to be carried as standards, adds strength to this concept of superfamilial groups. The groups in turn may have been part of a larger two-part grouping, as is suggested by the motif of the "Lord of the Beasts" on the one hand and the motif of the "Pipal-deity worshipers" on the other. Such a possible structure is familiar to anthropologists as what is termed a moiety: a society characterized by the classification of clans or similar subsidiary organizations into two groups, "halves" that usually intermarry.

If one accepts the working hypothesis that the pictorial material on each seal identifies its bearer as to clan and moiety, it logically follows that the part of the seal devoted to Harappan script could be concerned with identifying the bearer as an individual. The script might, for example, give the individual's name, occupation, place of residence, rank or title and similar information. In support of such an interpretation the study of seal texts reveals considerable variation in the sequence of individual signs and yet a frequent repetition of certain of them. The hypothesis gives the would-be decipherer some basis on which to proceed. Just as Michael Ventris knew that at least some of the Linear B texts found on Crete were inventories of material objects, so the decipherer of Harappan script can assume that what to search for are proper names and their embellishments or other statements identifying the individual.

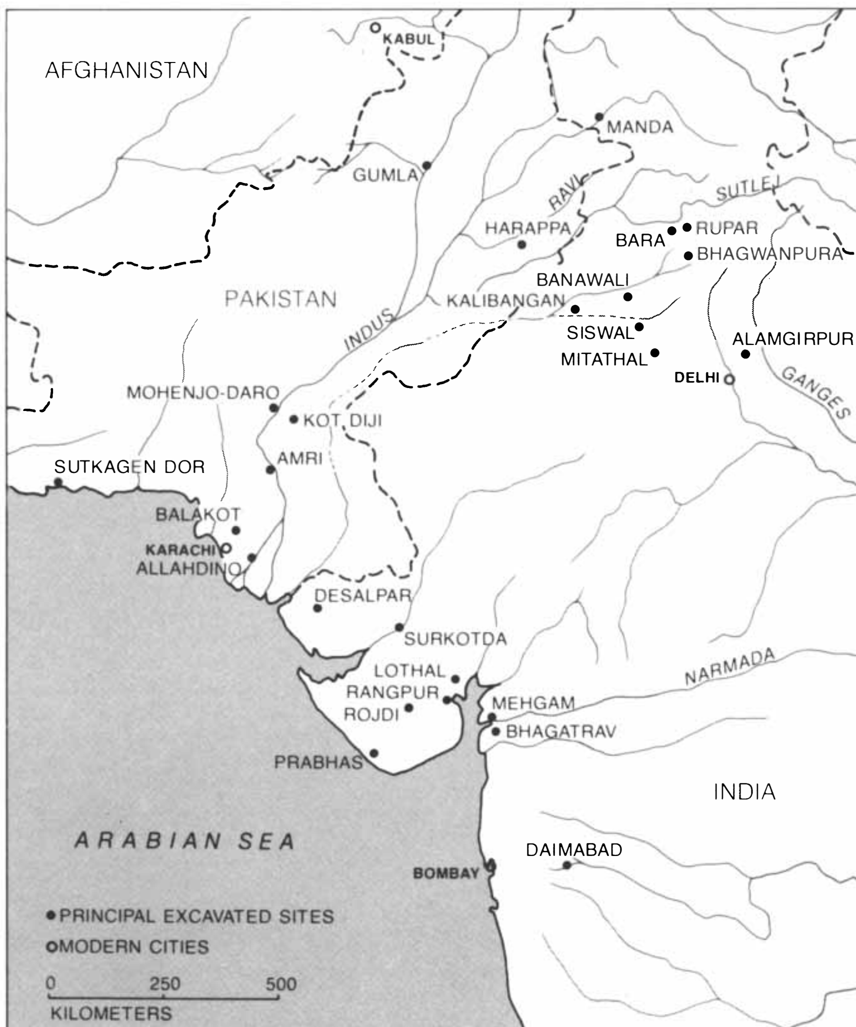
As long ago as the 1930's the British scholar G. R. Hunter had identified a total of 396 separate Harappan signs. More recent work has added another 23, so that the script is now known to incorporate a total of 419 signs. Statistical analysis shows that they occur 13,376 times in 2,290 known texts. Of the 419, 113 signs occur only once, 47 occur twice and 59 occur fewer than five times. In effect this means that the remaining 200 signs were in more or less general service, and analysis shows that fully half of them are combinations of the remaining half.

These findings demonstrate that Harappan writing was neither alphabetic, as Sanskrit is, nor logographic (that is, having one character for each word), as Chinese is. This places Harappan writing in the category known as logo-syllabic, meaning that some signs represent words and others serve purely for their syllabic values, or sounds. Other examples of this kind of writing are Egyptian hieroglyphs, early Sumerian ideographs and modern Japanese. The fact that half of the Harappan signs in common use were combinations of other commonly used ones suggests that the writers exploited such combinations to express ideas (as the Chinese do when they pair the characters for sun and moon to represent the word "brightness") and to combine syllabic sounds to "spell out" a word.

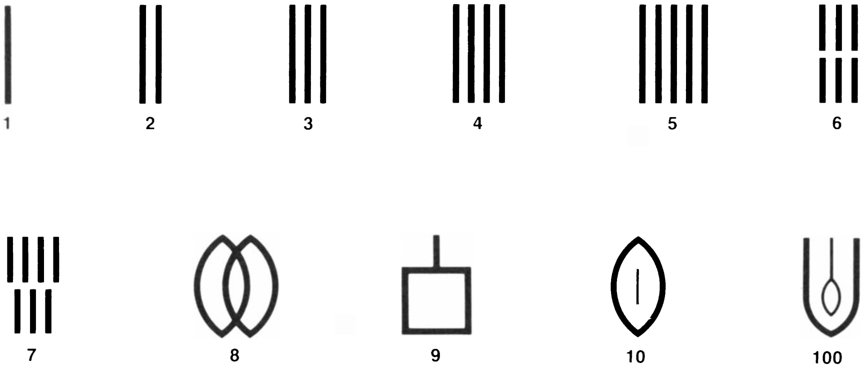
A crucial part of any writing system is the series of devices used to indicate gender, to distinguish between singular and plural, to establish the case of a verb and so on. The identification of these devices goes a long way toward establishing the relations between the graphemes, or individual components, of words and the language the graphemes represent. Now, the Harappan texts exhibit certain regularly paired signs, much as in the English reiteration of titles such as "His Majesty" and "Her Grace." It is also notable that certain signs appear in the middle of a text but rarely at the beginning or the end, whereas with other signs the reverse is true.

The usual ordering of signs and the identity of characteristic pairings may be established with a grid. Since the Harappan inscriptions are commonly brief, such a grid can consist of a relatively limited number of vertical columns. In fact, 14 columns were found to be sufficient. Of the first 17 inscriptions selected for horizontal grid display, most consisted of only five or six signs; thus they were entered near the center of the column array, allowing any longer lines to extend to the right or the left.

On entering the texts on the grid it became evident that certain signs appeared regularly in most of them. Con-



HARAPPAN SITES have been found from almost as far south as Bombay to as far north as central Asia and from as far east as Baluchistan to as far west as Delhi. Twenty-five of the principal excavated sites are shown on this map, including the great cities of the civilization: Mohenjo-Daro and Harappa. Most Harappan sites are small, covering between two and five acres.



HARAPPAN NUMERALS, perhaps originally part of a system with a base of eight, are a plausible set of vertical strokes for 1 through 5 and a sextuplet of short strokes for 6. The sign for 7 is a similar set of short strokes. The signs for 8 and 9 are respectively pictographs of a double sun and a "foundation post." The sign for 10 is a single sun with a short stroke inside it. One other identified numeral, the sign for 100, is a pictograph of a mortar and pestle.

sider the "pot" signs in column 5, the "loom twist" signs in column 8 and the "two-stroke" signs in column 10. When two of these three signs appear in the same text (as in lines *b, e, h* and *p*), they are always in the same right-to-left order with respect to each other no matter what other signs are included in the inscription. Regularity of position evidently governed their relations. Accordingly in setting up the Harappan grid such signs were placed in the columns noted above even when only one of them appeared in a particular text.

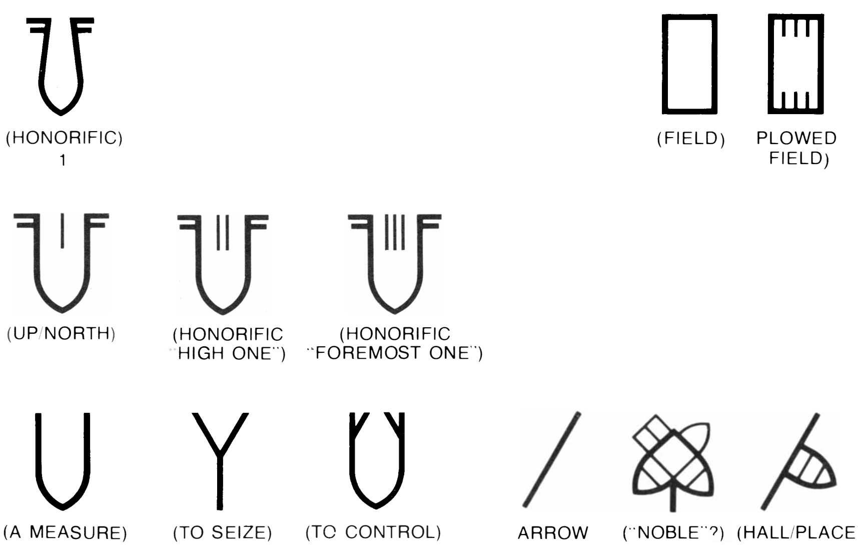
What then became clear was that these signs were the most numerous among those in the 17 selected texts. The pot sign in column 5 appears there 10 times, and three variants appear in two other columns. The loom-twist sign in column 8, variants included, appears six times, and the two short vertical strokes of column 10, variants included, appear eight times. Certain other signs utilizing vertical strokes are not as fixed in their relative positions. Whereas the group of one or two short single strokes can be accommodated in column 10, those groups consisting of one, two, three or more long strokes fall outside the central columns of the grid.

By now, acquainted with the appearance of some 50 Harappan signs as they are displayed on the analytical grid, the reader may be wondering in which horizontal direction they are to be "read." On the basis of seal inscriptions alone this might have been a hard question to answer. Fortunately the graffiti inscribed on pottery supplied an answer. Studies by B. B. Lal and I. Mahadevan of the Archaeological Survey of India have shown that some graffiti have overlapping strokes. The overlaps demonstrate that the direction of writing was from right to left. That is why column 14 of the grid appears at its left margin and column 1 at its right margin. The reversal puts the reading order in the more familiar left-to-right pattern.

We now come to the most complex part of the decipherment problem: What was the Harappan language? When Ventris identified the language of Linear B as Greek, he had overcome his greatest difficulty. The archaeological record of the region offers something of the same kind of help to those trying to pick a candidate for the unknown tongue. For example, the record shows that Harappan civilization was no sudden development. It had a long ancestry in the Indo-Iranian borderland, spread widely and eventually made its own contribution to the emergence of village India. This sequence suggests that the language spoken by the Harappans cannot have completely disappeared from the subcontinent.

Accepting such an assumption, which of the three principal families of languages spoken in the region might be related to Harappan? One candidate is Munda, a family of languages (spoken largely in eastern India) that seems affiliated with certain languages of Southeast Asia. Studies of the earliest forms of Munda, however, find little in its vocabulary that comports with what archaeology tells us about the culture of the Harappans. Another candidate is Indo-Aryan, a family of languages that traditionally came to India in the middle of the second millennium B.C. Its earliest literary expression, the Rig-Veda, however, describes a basically central-Asian culture quite different from the Harappan. The third candidate is Dravidian, a language now spoken mostly in southern and southeastern India but also still found in pockets in northern India and in Baluchistan.

A fourth possibility, of course, is that Harappan is related to none of the above. It is nonetheless worthwhile to consider the candidacy of Dravidian more closely than that of either Munda or Indo-Aryan. In addition to the northern forms (Kuruḵẖ, Malto and Brahui) about 25 Dravidian languages are still spoken. Indeed, the major families (Tamil, Malayalam, Kannaḍa and Telegu) are spoken by more than 100 million people. Thanks again to the archaeological record it is not necessary to ac-



COMBINED SIGNS make up half of the total of the 200 signs most frequently encountered in the Harappan script. Four examples are given here. At the top left is one of the commonest script signs, a pictograph of a pot with handles (labeled with the Arabic numeral 1). Below it are three combined signs: the sign labeled *1* plus signs like those for the Harappan numerals 1, 2 and 3 but consisting of short strokes. Under each of these three signs appears its sense in English. Below the combined signs are seen, from left to right, the pictographs for a container and for a forked stick, and a third sign combining them. At the top right is the pictograph for a field, a simple rectangle. Beside it is a sign that combines the field sign with the strokes of the Harappan numeral 6 to form a new pictograph with the apparent sense of "plowed field." At the bottom right, from left to right, are pictographs for "arrow" and for "pipal-tree leaf." The third sign, combining "arrow" with a partial "pipal leaf," seems to indicate "hall" or "place."

cept the candidacy of Dravidian blindly.

In 1974 excavations at the Harappan site of Allahdino, near Karachi, unearthed an ivory fragment with a half-circular cross section. There were holes on one side, apparently drilled to hold pegs; on the other side were two lengthwise parallel grooves. The object was similar to a large number of ivory rods or sticks excavated in the 1930's by one of the later workers at the great Harappan city of Mohenjo-Daro, E. J. H. Mackay. The Mohenjo-Daro ivories were not, as has been asserted, merely gaming pieces. For example, one of them, square in cross section, had a series of alternating circles and crescents cut into one side.

On some of the Mohenjo-Daro ivories is a sign seen in the grid (column 6, line *c*) and another sign resembling some form of plant. In various seal texts these two signs are associated with vertical strokes that range from one to seven in number and with five other signs. (Two appear on the grid at column 7, line *c*, and column 11, line *q*.) Like the second sign on the Mohenjo-Daro ivories, the sign at column 6, line *c*, appears to represent a plant, probably a stalk of grain. Considering the association of both signs in the seal inscriptions with what appear to be numbers and the marking of one of the ivories with circles and crescents that approximate in number a lunar (crescent) month of 30 solar (circle) days, the two plant signs might logically be taken to represent a word meaning both grain and month (or moon). Furthermore, the vertical strokes associated with both signs end after reaching a total of seven, suggesting that some other sign served for the number eight and possibly that the Harappan number system was to the base eight.

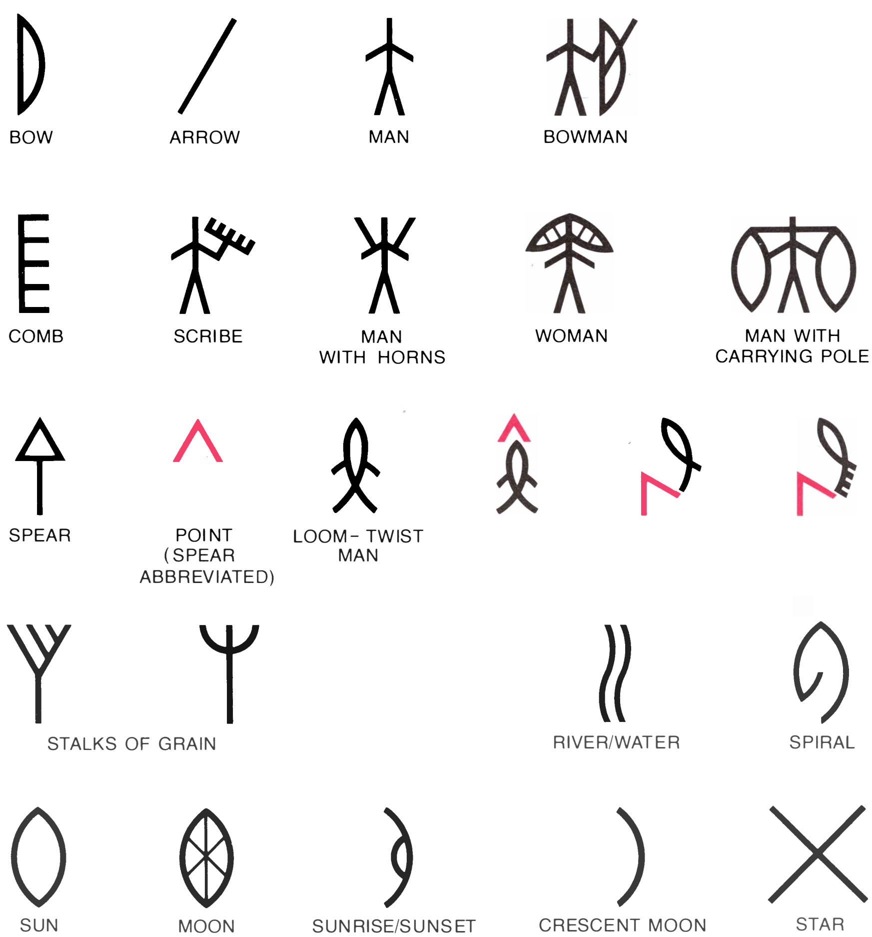
What language has a word for grain that also means month or moon and is associated with a base-eight numerical system? On the first point *nel* means rice in five Dravidian languages and *nilā* or *nela* means moon in three of the same five and in five others as well. It is also the word for month in some of the same languages. On the second point, a student of Dravidian language, Kamir V. Zvelebil, has pointed out that the original Dravidian number system was indeed probably to the base eight: the count to 10, used for conformity with the base-10 number system today, goes literally (in translation) "one," "two," "three," "four," "five," "six," "seven," "number," "many minus one" and "many."

Over the years scholars have noted in Dravidian a number of homophones: words with the same sounds but with different meanings. For example, the word for the shoulder pole from which pots are suspended is *kā*. It is also the word meaning guardian, or protector. The common word for fish is *mīn*, which is also the word for star. The word for 100 is *nūṟu*, which is also the word meaning to grind or to powder.

The homophonic, or rebus, principle is found in a number of ancient languages, including the hieroglyphs of Egypt. The key aspect of the homophonic principle is that it seeks a syllabic equivalent (say a picture of an eye to mean "I") rather than merely being a picture of something. For example, an early Egyptian ruler, Nr-mr, was represented in hieroglyphs by the sign for a catfish (*nr*) and the sign for a chisel (*mr*). The representation was not meant to suggest that he be called "the catfish-chisel one" but rather that his name sounded like "catfish-chisel."

Does this suggest a basis for deriving a Harappan syllabary? I shall put the point to trial in what follows, first warning that the acid test of any decipherment is its consistency. If a decision is made to declare a certain symbol equivalent to a selected sound or sense, it cannot later be shifted to represent some other sound or sense. Inner logic is basic to all writing systems; when decipherers arbitrarily alter the values of symbols in order to fit the model they prefer, their work is spoiled.

For a start consider the sign shown in column 7, line *h*. Pictographically it could be taken to represent a mortar and pestle. Linguistic reconstructions suggest that this utensil was given the syllabic value *nūṟu* in Dravidian. This, as we have seen, is also the sound of the verb to grind or powder and the numerical noun "100." Next consider the sign that figures so prominently in column 5 of the grid. Some years ago Mahadevan conjectured that this sign, which most



ADDITIONAL PICTOGRAPHS to be found on the grid on the next two pages include those illustrated here together with some combined forms. In the top row, from left to right, are "bow," "arrow," "man" and the combination "bowman." In the second row are "comb," a third human stick figure holding a comb (a combination that appears to indicate "scribe"), a fourth stick figure with "horns," a fifth stick figure with an elaborate coiffure ("woman") and a stick figure with a loaded shoulder pole (a combination appearing to indicate "guardian"). In the third row, from left to right, are "spear" and the shorthand for "spear," "point" (*color*), an honorific; a "loom twist" plus human arms, a second honorific pertaining to rulership, and the two signs combined. In the fourth row are two stalks of grain (both signs have the interchangeable sense of "grain" or "month"), a sign with the sense of "river" or "water" and a spiral with the sense of "surround" or "surrounding." In the bottom row are five astronomical pictographs. From left to right are seen the sun, the moon, sunrise or sunset, the crescent moon and a star.

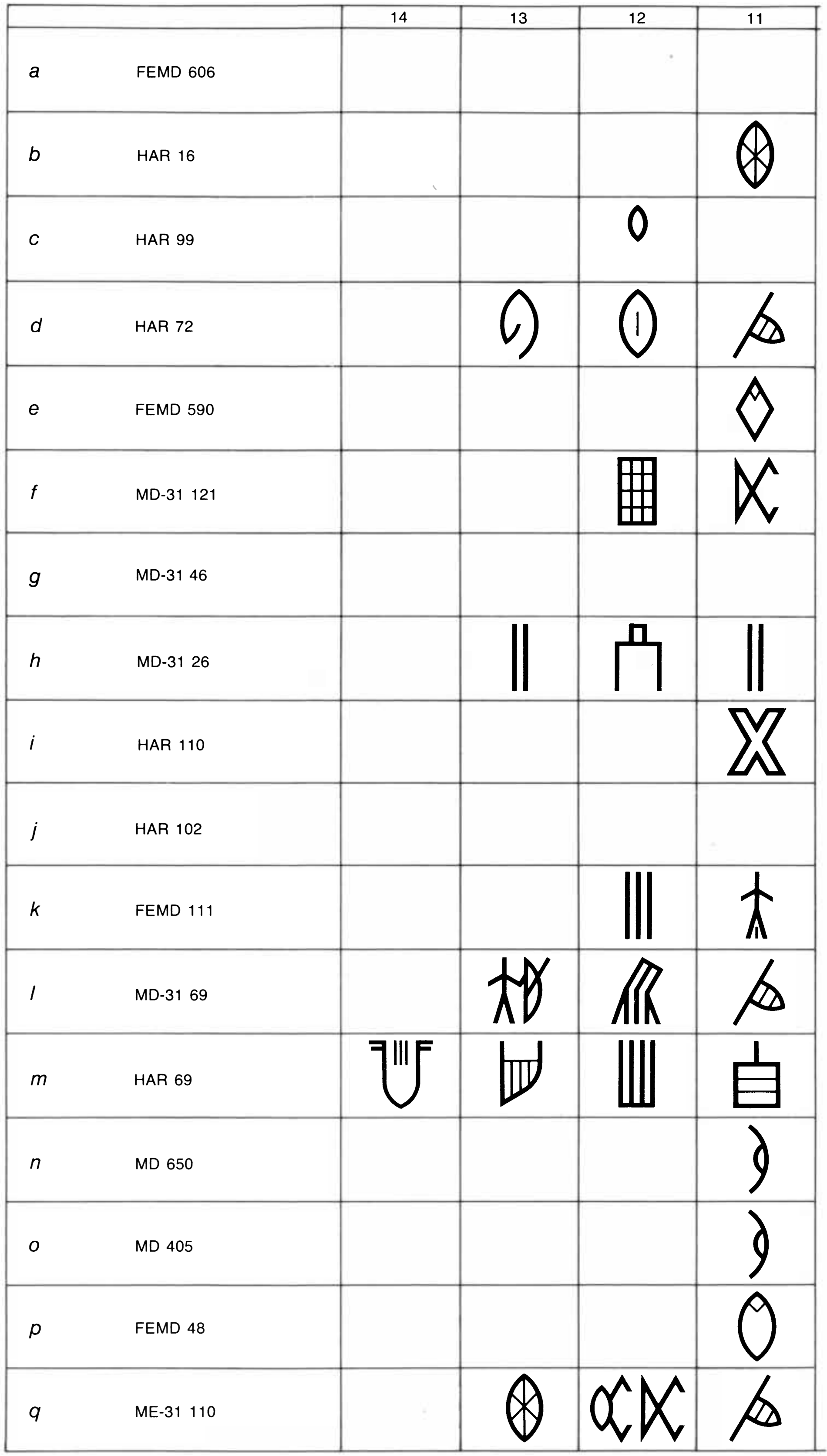
frequently appears at the end of seal texts, was a pictographic representation of a pot with handles. He went on to point out that various Dravidian words for this kind of pot were homophones of words meaning male, including the common honorific suffix *an̲* used with male personal names at least as long ago as early in the Christian Era. The fact that such a sign terminates seal texts hypothesized to contain personal names strengthens Mahadevan's conjecture.

Now, another terminal sign on the grid (column 5, line *f*) is a pictograph of a human figure carrying a shoulder pole with a pot at each end. As we have seen, the syllable for a pole with pots alone, *kā,* is a homophone of the Dravidian words meaning "to guard" or "to protect." Furthermore, a common Dravidian word for man is *āḷ.* The combination of these would allow the reconstruction of the sign at column 5, line *f,* as a two-syllable word, *kā-āḷ.* In the Dravidian languages, for the sake of euphony, such adjacent vowels are separated by a consonant, either a *v* or a *y.* Thus the restored word, *kāvāḷ,* could be translated as "one who guards or protects," a statement of personal identity suitable for entry on a private seal.

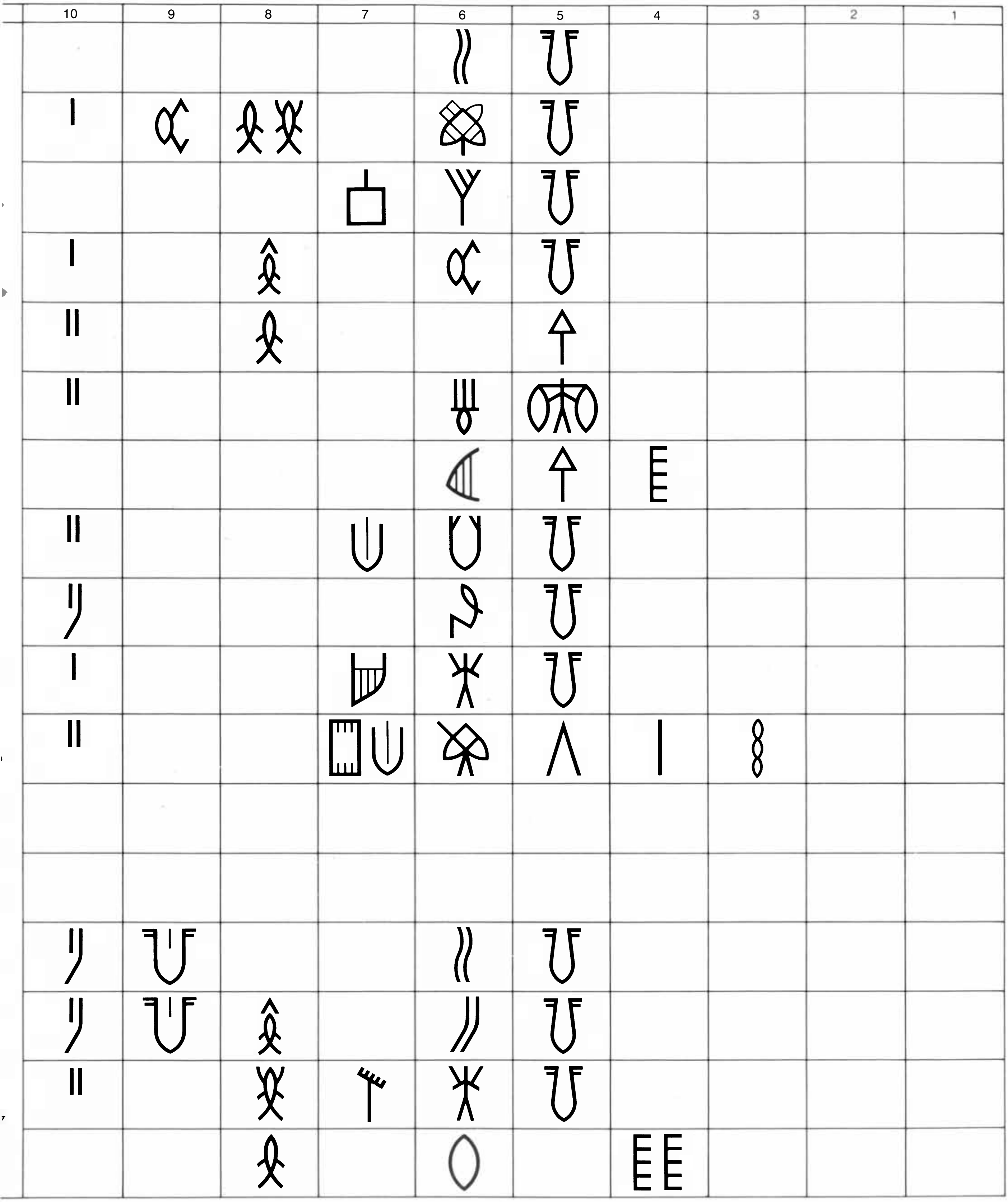
To give two further examples before turning to trial readings of some of the seal texts on the grid, two of the "number" signs—the sign in column 12, line *k,* presumably meaning three, and a second sign consisting of four vertical strokes, presumably meaning four—also have plausible homophones. The first homophone, *mu*(*n*), equates with the Dravidian word meaning "foremost" (*mun̲*); the second homophone, *nāl,* equates with the word meaning "good" (*nal*), both logical titular adjectives.

Let us now consider two short and two longer seal texts from the grid. The text designated Mohenjo-Daro 31 46 (line *g*) consists of three signs, putatively a musical instrument, a spear and a comb. The assigned Dravidian syllabic values are, in the same order, *pan, ār* and *ki*(*r*), in translation the word "sing," a third-person honorific and the word "mark," read as *Panār-ki*(*r*). The proposed translation is "Panar's seal," or literally "The singer's mark." Next, the text designated FEMD 590 (line *e*) consists of four signs. The first sign, a diamond shape, has the three-syllable value *āra-man,* the second, two short vertical lines, has the value *i*(*l*), the third, a "loom twist," has the three-syllable value *piri*(*key*) and the fourth is the same honorific *ār* as in the preceding text. This is read as *Āra-man-i*(*l*) *pirikeyār,* with the proposed translation "Belonging to the noble house (of the) Pirikeyar" ("pirs" and "pirikeys" are chiefs).

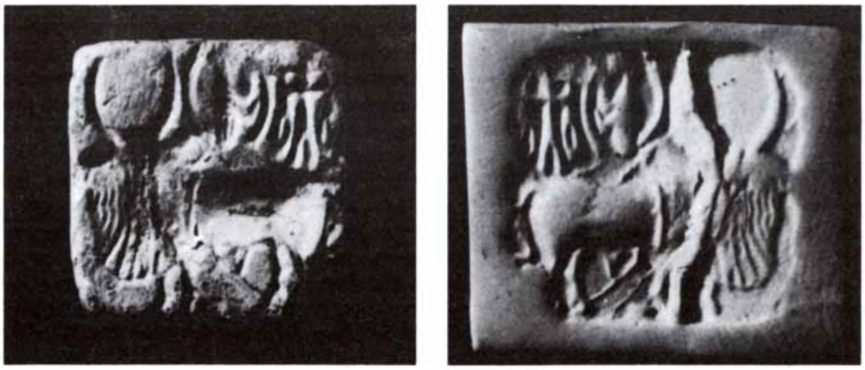
The third text, Harappa 72 (line *d*), consists of seven signs: an initial spiral (*cur*), an oval enclosing a short vertical



ANALYTICAL GRID 14 columns wide contains 17 seal texts that range in length from two signs to nine. When "pot" signs were all assigned to column 5, "loom twist" signs and their variants to column 8 and the signs made up of two short vertical dashes to column 10, it became apparent that when two or more of these three signs appeared in an inscription, they always appeared in the same right-to-left order with respect to one another regardless of what other signs might appear in the text. Lines *b, e, h* and *p* provide four examples. The single short verti-



cal dashes in column 10 (lines *b, d* and *j*) reveal an inflectional quality. The same sign also appears in combining form in column 12, line *d*, as an addition to a "sun" sign; in column 11, line *k*, as an addition to a "man" sign, and in column 9, lines *n* and *o*, as additions to "pot" signs. The sense of the single sign and its combination (lines *i, n* and *o*) appears to be that of a possessive; the sense of the double-dash sign appears to be that of a locative. Two of the texts, lines *l* and *m*, do not fall within the "normal" central grid distribution. Thus it can be assumed that they are not texts of the usual formulation. When two signs are found in one grid cell, the placement indicates that the signs are regularly found paired in seal texts. Readings of four of these 17 texts (lines *d, e, g* and *k*) are shown at the bottom of the next page.



HARAPPAN CONTACT WITH MESOPOTAMIA is evidenced by the discovery of Harappan seals in the region. The one shown here, with the seal at the left and the sealing impression at the right, was uncovered in 1975 in the ruins of a house of the Kassite period at Nippur, an ancient city in Iraq, by McGuire Gibson of the Oriental Institute of the University of Chicago.

dash (*patǎ*), a pipal leaf and an arrow combined (*ambara*), a short vertical dash alone (*ǎ*), a twist with a point above it (*āra-pirikey*), pincers combined with the sign for the sun (*paṭu-kāru*) and, as the terminal sign, the pot with handles (*aṉ*). This is a mouthful: *Cur patǎmbara-ǎ āra-pirikey patu-kāran*. The proposed translation is "Patukaran, powerful (noble) chief of the surrounding settlements." The fourth text, FEMD 111 (line *k*), is one of the longest of the translated seal inscriptions, consisting of nine signs. In the interest of brevity, only the syllables and their proposed translation will be given: *Munālǎ-i(l) nūṟu cāṟuvara amban ārōru malya,* or "Belonging to Munala, mistress of 100 plowed fields, noble first lady."

So far syllabic values have been assigned to nearly 100 seal and pottery signs and appropriate Dravidian homophones have been found for each. Translations such as those given here have been proposed for more than 100 seal texts. They range from such simple statements as "water-holder" (probably a proper name) to such sonorities as "Arasamban, High Chief (of) Chiefs of the Southwest, lineage of the Moon."

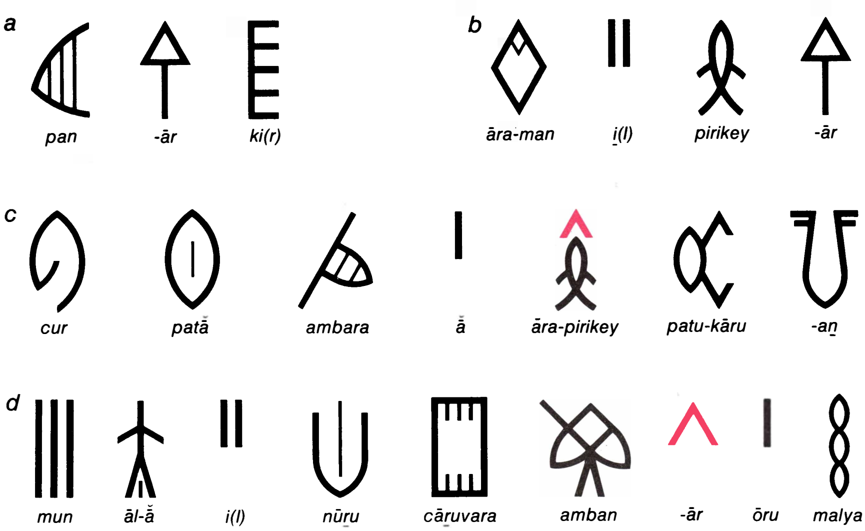
What such readings demonstrate about the Harappans is that a number of individuals (Arasamban among them) traced their lineage to such major figures of the cosmos as the Sun, the Moon and the Stars and perhaps also the Monsoon Rain. These may represent distinctions within each clan. Chiefs were associated with *aramani,* or chiefs' houses, which may have had both residential and administrative functions. In addition there was a high place (the "citadels" identified by archaeologists?) that served a special but so far undetermined function. An assembly area—an open court or a pillared hall (both known archaeologically)—suggests that an assembly of chiefs was a basic part of Harappan political organization.

Among other kinds of leaders known from the seals were heads of associations (guilds?) such as that of the coppersmiths, storehouse overseers, irrigation supervisors and landowners, a category that as we have seen included women. Religious references, however, are scarce. There is a possible "horned deity," referred to as "the copper one" or "the red one," and a possible "mother deity," but at the moment these distinctions are not confirmed. Lesser figures included drummers and singers, the drummers perhaps to summon assemblies and the singers perhaps to entertain or perform at ceremonies.

The seals attest further to a class of scribes, to people in charge of weights and measures and to supervisors of the distribution of stores, the grinding of flour and probably hunting operations. There were also captains of boats and custodians of fire. Many seals bear the sign *kā,* referring to guardianship, not in a military sense but more in terms of a responsibility for the care of crops and the preservation of herds and flocks. One of the gratifying near-proofs that the decipherments are on the right track comes from the prismatic seal where a gavial is twice seen above an array of "clan" animals. In Dravidian the word for crocodile is *mutalai.* The Dravidian word for first chief, *mutali,* is a close homophone.

Thus it appears that the language of Mohenjo-Daro and Harappa some 4,000 years ago was an early Dravidian tongue and that the Harappan scribes struggled to put that language into graphic form as a method of identifying the elite of the Indus Valley civilization. The Harappan civilization, however, was geographically so widespread that a number of non-Dravidian words must have entered the language, just as Dravidian words were later borrowed by the speakers of Indo-Aryan.

A great deal of additional decipherment remains to be done and no doubt what has been put forth here will be found to have flaws. What remains to be done is even more exciting. It leads toward a goal that until recently seemed impossible to attain: cohesive information on Harappan polity, social organization and ideology, and perhaps even stronger evidence that, as the parent of the village India of today, the Harappan culture never did disappear.



FOUR SEAL INSCRIPTIONS appear together with the syllabic value of each sign in Dravidian, a surviving family of early languages of the region. The shortest inscription (*a*), in the author's reading, is "The singer's mark." The next inscription (*b*), which repeats the third-person honorific "spear" sign, *ār,* is read as "Belonging to the noble house (of the) Pirikeyar." The next inscription (*c*) is seven signs long and includes a variant of the "loom twist" sign of inscription *b*. It is read as "Patukaran, powerful (noble) chief of the surrounding settlements." The last inscription (*d*) is nine signs long and includes the two-stroke sign *i(l)* of inscription *b* and an abbreviation of the "spear" signs in both *a* and *b* (*color*). It is read as "Belonging to Munala, mistress of 100 plowed fields, noble first lady." More than 100 seal texts have now been read. The majority of the readings accord with the hypothesis that the seal signs identify individuals.

We discovered a new kind of brain cell.

You'll find it at places like the Institute for Advanced Study, where experts from a dozen different disciplines cross-pollinate their thoughts and ideas.

But as far as companies in America go, we at Kodak may be the only ones who can bring together our own experts in programming, optics, physics, organic and polymer chemistry, fibers, human engineering, imaging, marketing, paper technology ...we've probably skipped a few.

Call it synergism or whatever. But when these pros sit down at our table, they come up with some truly incredible results.

The Kodak Ektaprint 250 duplicator is a good case in point. It has grown out of more disciplines than any single Kodak product before it.

The nucleus is two 8085 microprocessors, with a combined yield of over 100,000 bytes, and responsibility for 134 separate functions.

There are all kinds of thoughtful touches. One is a programmable display that's as friendly as can be—with 94 different messages, not codes, to help anticipate and satisfy operator needs. And there's an internal diagnostic system that's amazingly specific.

Then there's single-pass duplexing. Lots of heads went together for this run-through-once, print-both-sides convenience. Collating and cover inserting, too.

A couple of other neat feats: Fiber specialists from our Eastman Chemicals Division in Tennessee helped create the "perfect" toner brush. In Rochester, our experts in 35 mm added sprockets to the belt drive. And a reflective "bathtub" was perfected by our processing group; it helps keep the paper path shorter and virtually trouble-free.

With all these new capabilities, you may think the 250 needs a genius to run it. Wrong. And that's the beauty. 99% of the duplicator is microprocessor-controlled. That means reliability, flexibility, productivity, easier operation. And time to think about more important matters.

For more information on how we plan to combine our talents for future applications in microelectronics, please send for our brochure ..."Resources To Innovate, Commitment To Serve..." (B3-190). Write to Eastman Kodak Company, Dept. GB-SA 6, 343 State Street, Rochester, New York 14650.

Kodak

Kodak. Where technology anticipates need.

SCIENCE AND THE CITIZEN

Under the Volcano

It has long been hypothesized that large volcanic eruptions can affect the climate worldwide for several years. By putting dust and ash into the atmosphere volcanoes are assumed to reduce the amount of sunlight reaching the earth's surface. The result could be a global cooling, although the mechanism has not been precisely determined. In part because of this imprecision the explosion last year of the Mexican volcano El Chichón has aroused considerable interest among geologists and meteorologists. Now Michael R. Rampino of the National Aeronautics and Space Administration's Goddard Institute for Space Studies and Stephen Self of Arizona State University suggest that the effect of El Chichón on global climate could be as great as that of much larger eruptions, which lowered mean annual temperatures in the Northern Hemisphere by about .4 degree Celsius.

El Chichón, which until the eruption was about 1,400 meters high, is in Chiapas, the southernmost state of Mexico. Chiapas is at the intersection of three major tectonic plates: the North American plate, the Caribbean plate and the Cocos plate, which lies under the Pacific Ocean off Central America. The subduction of the Cocos plate under Mexico is thought to be the cause of volcanic activity in a broad belt in southern Mexico; El Chichón lies at the southern end of the zone. It had been dormant throughout history until it erupted with great force on March 28 and again on April 3 and April 7.

Almost immediately after each of the major eruptions a large cloud of dust, ash and gas appeared in satellite photographs of Mexico. The three clouds, reaching a height of about 16 kilometers, soon merged into a single mass. The lower part of the combined cloud moved northeast toward the Gulf of Mexico; it was quickly thinned out by the settling of dust particles and the scavenging action of ordinary water clouds. The higher part of the cloud, extending into the stratosphere, was driven westward over the Pacific.

In mid-April the stratospheric mass was detected over Hawaii by measurements of the amount of solar radiation reaching the earth. In May the cloud was observed by workers in Ahmadabad, India, who reported that it extended from an altitude of 17 kilometers to 30 kilometers, a range that includes most of the stratosphere. It is thought the stratospheric cloud circled the earth in a narrow belt, then diffused to higher and lower latitudes and eventually covered the globe completely.

To what extent will the El Chichón cloud affect the weather? Writing in *Quaternary Research,* Rampino and Self have given a basis for answering the question by correlating the amount and the type of matter ejected in three big volcanic eruptions with subsequent changes in mean annual temperature in the Northern Hemisphere. Tambora, on Sumbawa Island in Indonesia, erupted in 1815 and caused a cooling of about .8 degree C. Krakatoa, in the Sunda Straits between Java and Sumatra, erupted in 1883 and reduced the average temperature by about .4 degree. Mount Agung, on Bali, erupted in 1963; again the temperature reduction was about .4 degree.

Most attempts to explain the effect of volcanoes on climate have focused on the total volume of matter ejected during the blast. Even though the Tambora, Krakatoa and Mount Agung eruptions had similar effects on climate, the volumes of rock ejected were very different: about 150 cubic kilometers for Tambora, 20 cubic kilometers for Krakatoa and one cubic kilometer for Mount Agung. In explaining the disproportionate effect of the eruption of Mount Agung, Rampino and Self cite the andesitic composition of the volcano. (Andesite is a volcanic rock intermediate in composition between the basaltic and the silica-rich types.) Andesitic rock tends to have more sulfur than the silica-rich rock of Tambora and Krakatoa. Analysis of rock samples has shown that the sulfur content of Mount Agung was 800 parts per million, compared with 380 for Tambora and 150 for Krakatoa.

In an eruption the sulfur is converted into sulfur dioxide gas, which reacts with water vapor in the stratosphere to yield a fine aerosol mist of sulfuric acid. Rampino and Self note that there seems to be a limit to how much acidic aerosol can be formed in the stratosphere; for this reason a large increase in the amount of ejected matter does not cause a corresponding increase in the amount of acid. Furthermore, above an altitude of 30 kilometers the droplets evaporate quickly; thus an exceptionally powerful eruption, which throws material high into the atmosphere, may yield a smaller quantity of aerosol than a less powerful one.

Rampino and Self estimated the amount of sulfur-bearing aerosol put into the atmosphere by each of the volcanoes by measuring the amount of sulfate deposited in the annual layers of the polar ice cap. The measurement showed that the aerosol yield from Tambora, Krakatoa and Mount Agung was in the ratio of 7.5 : 3 : 1, much smaller than the ratio of the total volume of ejected matter. Moreover, given the imprecision of the estimates, the sulfate ratio is close to the observed ratio of temperature reductions. Rampino and Self conclude that it is the acidic aerosol that makes the largest contribution to blocking solar radiation.

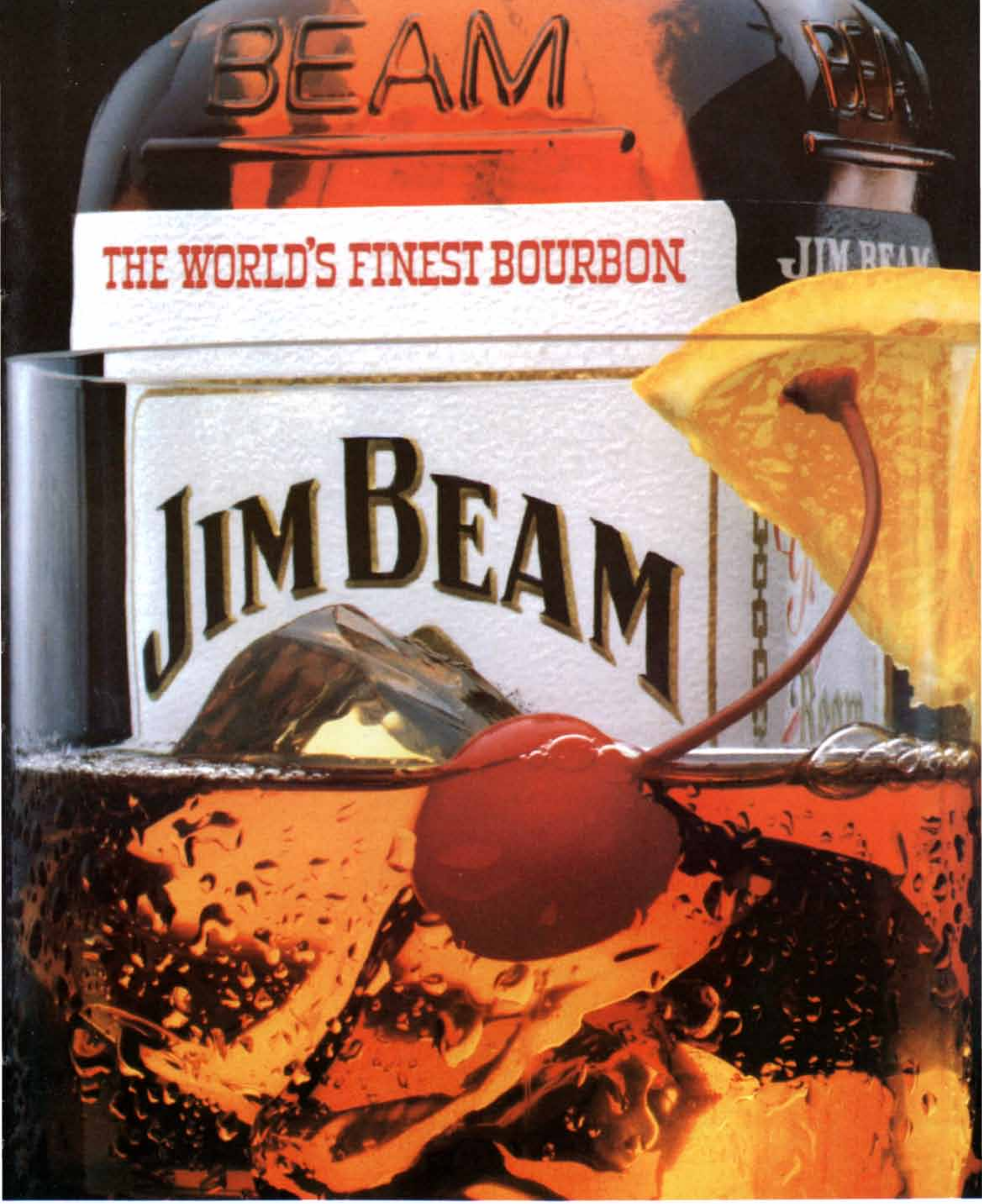
The eruption of El Chichón was small compared with the three historical examples: it is estimated to have ejected about .5 cubic kilometer of rock. Like Mount Agung, however, El Chichón is an andesitic volcano. Furthermore, it is "exceptionally rich in sulfur," even richer than the rock type would indicate, Rampino has said. "There may have been an input of sulfur from sulfates in rock below the volcano," he added. Samples taken near the volcano have suggested there is a stratum laden with calcium sulfate.

Because of El Chichón's sulfur content and because its eruption had the force needed to blanket the stratosphere, it may have an effect on climate as great as that of Mount Agung, Rampino said. The mean temperature in the Northern Hemisphere could be lowered by as much as .5 degree this year and next. It should be noted, however, that although such a cooling is measurable, it is of roughly the same magnitude as random year-to-year fluctuations.

Rapid Pulse

The study of pulsars, compact sources of periodic radio emission in the galaxy, has been disarrayed at least temporarily by the discovery of the fastest one yet: a pulsar in the constellation Vulpecula pulsing so rapidly that it may prove to represent a new class of astrophysical objects. When the newly discovered pulsar first came to attention, it was hidden in a region of radio emission designated 4C 21.53. Then the pulsar itself was identified and was designated 1937 + 214. Investigators who are trying to account for its extraordinary properties suspect it has had an eventful billion-year history. They call it a recycled pulsar.

What made 1937 + 214 stand out in the radio sky were two attributes of 4C 21.53. First, the radio spectrum of 4C 21.53 was known to decrease steeply with increasing frequency, a characteristic of pulsars. Second, 4C 21.53 appeared to twinkle at radio frequencies, much as the image of a star twinkles at optical frequencies. The optical twinkling is caused by irregularities in the atmosphere of the earth; the radio twinkling is caused by analogous irregularities in the solar wind. A source of radiation can twinkle only if it is smaller than a typical irregularity, and so the radio source in 4C 21.53 could have an angu-


BEAM
THE WORLD'S FINEST BOURBON
JIM BEAM

lar diameter no larger than one second of arc, or .0006 times the diameter of the sun. That too suggested a pulsar.

Yet no pulsar had been detected in 4C 21.53. Examining this puzzle in 1979, D. C. Backer of the University of California at Berkeley decided the reason might well be a phenomenon called pulse broadening. Here the radiation from a pulsar is disrupted by turbulence in the ionized gas that permeates interstellar space. The parts of a pulse that are not disrupted reach the earth with the least delay; the parts that follow zigzag trajectories take longer. The effect is to smear out the pulses, particularly if the pulses are short. Backer estimated that if a pulsar in 4C 21.53 emits pulses with a period of less than 10 milliseconds, it would not have been detected in the searches made up to then. This limit, however, was one-third the period of the fastest pulsar known.

In 1979 and again several times last year Backer and his colleagues undertook further searches. They examined radio wavelengths of about 20 centimeters, where the disruption caused by the interstellar medium is minimal. Last November a search succeeded. Backer and his colleagues found 1937 + 214 in signals detected by the giant radio telescope of the Arecibo Observatory in Puerto Rico. Its pulses have a period of 1.5578 milliseconds, and so it pulses 20 times faster than the fastest pulsar known before.

Pulsars are taken to be spinning neutron stars that broadcast electromagnetic radiation in a beam, like a lighthouse. Thus their pulse rate equals their spin rate. It follows that 1937 + 214 spins once every 1.5578 milliseconds, or 642 times per second. The rate is alarmingly fast. Suppose 1937 + 214 has the mass of the sun and a radius of 10 kilometers (values thought to be typical for a neutron star). Then at its equator the surface of the star is moving at 13 percent of the speed of light. Moreover, the rotational energy of the body amounts to 7×10^{51} ergs. "This energy," Backer and his colleagues write in *Nature,* "is comparable to the entire mechanical energy output of a supernova," the stellar explosion that it is thought can give rise to a neutron star. The rotation rate approaches a natural limit: if it were about three times as fast, centrifugal force would overcome the gravitational self-adhesion of the pulsar and it would fly apart.

All pulsars slow down as they lose energy, and in general the deceleration is greater for pulsars that are spinning faster. The rate at which 1937 + 214 is slowing, however, is the lowest ever observed. In November it was established that the pulse rate (and hence the spin rate) is decreasing by no more than 10^{-15} second per second. Even before a recent measurement by Backer and his colleagues, which indicates that the spin of 1937 + 214 is decreasing by only 10^{-19} second per second, it had become questionable whether the accepted hypothesis of how a pulsar is formed applies to the newly discovered one.

In the accepted hypothesis a supernova explosion leaves behind a stellar remnant (a neutron star) with a high rate of spin and an intense magnetic field. In effect the tiny neutron star preserves the parent star's angular momentum and magnetic flux. The neutron star loses energy in part because a spinning dipolar magnetic field (in essence a spinning bar magnet) can emit electromagnetic waves and in part because a spinning body that is not a perfect sphere is thought to give off gravity waves. When the neutron star is young and spinning fast, the rate of loss is great; therefore the rotation rate decreases and (in the view of some theorists) the magnetic field subsides.

How can 1937 + 214 be spinning fast and yet be losing energy slowly? In hypotheses advanced by several investigators, notably M. A. Alpar of Columbia University and his colleagues there and at Rutgers University, the answer is that the pulsar has been recycled. In one scheme 1937 + 214 was created in a binary star system when a supernova explosion left behind a neutron star. The companion star was undisturbed. The neutron star followed the normal evolution of a pulsar: it spun rapidly and lost energy, so that its spinning slowed and its magnetic field was dissipated. It may even have been extinguished as a source of radio pulses. Over billions of years, however, the companion star lost mass. Each increment of mass that passed from it to the neutron star brought with it an increment of angular momentum. The neutron star was thereby "spun up" again. Today its rotational energy is great but its magnetic field is small, so that it loses little energy. At some point the companion star was destroyed in a second supernova explosion, which left behind no solid remnant. By now even the gaseous debris has dispersed. Only the recycled pulsar remains.

Social Aging

On being cast out of Eden mankind was condemned to a life of toil and to the pain of childbearing. Ever since it has been the norm that almost all the years of adulthood are spent working and raising children. The pattern is now changing in the developed countries, but not everyone will welcome the change as a return to the Garden. Many people, after all, count work and children among life's satisfactions rather than among its burdens.

The causes of the change are demographic. The fraction of a population that survives to the threshold of old age has increased dramatically in the past 100 years; the life expectancy of an individual who has reached age 60 has also increased, although to a lesser extent. Furthermore, people are having fewer children, and the age at which they have their first child decreased through most of the 20th century; so has the age at retirement. As a result the period spent "at liberty," with neither parental responsibilities nor an obligation to work, has been greatly lengthened.

Massimo Livi-Bacci of the University of Florence discusses these trends in *Population and Development Review* and presents a quantitative analysis of their effects on the population of Italy from 1881 through 1981. The analysis is made somewhat uncertain by the limitations of the available data, which are taken mainly from census reports. For example, the Italian censuses have not recorded the age of parents when their last child is born, which determines the parents' age when all their children have left home. Nevertheless, the missing data can be plausibly estimated from the known age distribution of the population and the fertility rate.

Longevity has of course improved substantially. In 1881 the average life expectancy at age 35 was less than 30 years; in 1981 a man of 35 could expect to live another 37 years and a woman another 43 years. Livi-Bacci shows, however, that the number of years spent without dependent children or employment has increased by a much larger factor. "For men these years were an insignificant portion of life in 1881 (only one year), but they have gradually increased to 12 years in 1981, or one-third of [the expectation of life at age 35]. For females the increase has been extraordinary: from five years in 1881 to almost 23 in 1981, or more than half the expectation of life at age 35."

The result, Livi-Bacci notes, is a postponement of biological aging but an earlier onset of social aging, "the process of relinquishing meaningful social functions, with the ensuing increased risk, for each individual, of becoming prematurely obsolete, senescent and estranged from society. It is in the increasing number of years... liberated from the two fundamental functions of parental responsibilities and work that social aging makes its mark."

It seems unlikely that the workless and childless period at the end of a typical life will ever grow much longer than it is now. Indeed, natural demographic adjustments may shift some of the years of reduced responsibility to an earlier stage of life. "Youth is a very busy period in the life cycle," Livi-Bacci writes. "Education is a long process; leisure is attractive and varied; work is difficult to find but part-time jobs are easily and frequently taken and left. Thus responsible childbearing could be postponed to

THE SEDAN THAT DEFINED PERFORMANCE FOR A GENERATION OF CARS UNFAMILIAR WITH THE TERM.

"In the beginning," wrote Car and Driver, "when BMW said, Let there be sports sedans, there were many other companies that failed to grasp the importance of this edict."

What went ungrasped was the nature of performance itself. Which is not just road-holding prowess, or agility through corners, or acceleration, but all of these in tandem.

And that is also an excellent definition of the BMW 320i.

PERFORMANCE: THE UNABRIDGED VERSION.

Like all BMW's, the 320i isn't so much the sum of its parts as a perfect balance of them.

Its fuel-injected 4-cylinder engine is the basis for BMW racing engines.

Its gearbox is, as one critic observed, "the perfect complement to the (320i's) willing engine."

Which is complemented by a suspension that removes the treachery from even serpentine stretches of roadway.

The result, according to Car and Driver: "maybe the best-balanced small sedan around."

VALUE, ACCORDING TO THE 320i.

What is a car really worth? Whatever the market will pay. By that standard, the 320i is valuable indeed. According to the NADA Used-Car Guide, 320i's have traditionally enjoyed one of the highest resale values of any car in their price class.

It's also one of the best protected, with a 3-year/36,000-mile limited warranty and a 6-year limited warranty against rust perforation.*

All things considered, it's understandable why other car makers "are already boasting that their next generation of sporty coupes will feel and drive like their target car, the 320i" (Motor Trend).

If you're unwilling to wait that long for someone else's interpretation of a 320i, we invite you to test drive the original at your nearest BMW dealer.

THE ULTIMATE DRIVING MACHINE.

*Warranty applies to automobiles purchased from authorized U.S. BMW dealers only. See your BMW dealer for details. European Tourist Delivery can be arranged through your authorized U.S. BMW dealer.

a later stage in the life cycle (the trend has perhaps already been initiated)."

LEP and SLC

In the effort to accelerate particles of matter to ever higher energies one accelerator after another has been designed, built, exploited and then supplanted. The linear accelerator was followed by the circular design, which in turn was followed by the storage ring, where beams of particles collide head on. What may well be the last and biggest of the storage rings is about to be constructed. At the same time plans are under way to build a linear colliding-beam device, which may turn out to be the prototype of a new generation of accelerators.

The new circular machine is called LEP, or the Large Electron-Positron collider; it will be built by the European Organization for Nuclear Research (CERN) on a site straddling the French-Swiss border near Geneva. The storage ring will be in a tunnel 27 kilometers in circumference bored under the plain between Geneva and the Jura Mountains and under the mountains themselves. The particles in the ring will reach energies of about 50 billion electron volts (GeV) and will complete a circuit of the loop in less than a tenth of a millisecond. The other project is the Stanford Linear Collider (SLC), which may be added to the facilities of the Stanford Linear Accelerator Center (SLAC). It too would have a beam energy of about 50 GeV. LEP is scheduled for completion in 1987. The SLC will be completed in 1986, if it receives funding in fiscal year 1984.

The aim of a particle accelerator is to convert kinetic energy into mass and thereby create new forms of matter. Most physicists now classify the fundamental constituents of matter in three "generations" of the particles called quarks and leptons. All the members of the first two generations have been identified in experiments, but two particles in the third generation have not yet been seen: the quark called top and the lepton called the tau-type neutrino. They are among the particles that will be sought in the new colliding-beam laboratories. Also on the physicists' most-wanted list are certain particles that transmit forces between the quarks and leptons, notably the particle called the neutral weak boson, or Z^0.

In a modern linear accelerator electrically charged particles are injected into a long evacuated tube and are accelerated along the tube by electromagnetic fields created in hundreds or thousands of accelerator cavities. The particles slam into a stationary target at the end of the tube. The limit to the beam energy a linear accelerator can reach is set by the cost of the cavities and their associated radio-frequency power supplies.

It was recognized early on that higher energies might be reached at lower cost by building a circular accelerator with only a few cavities, through which the particles would pass many times. The particles are kept on a circular trajectory by bending magnets interspersed with the accelerator cavities. Ultimately the beam is extracted from the ring and directed toward a target. The commonest circular accelerator today is the synchrotron.

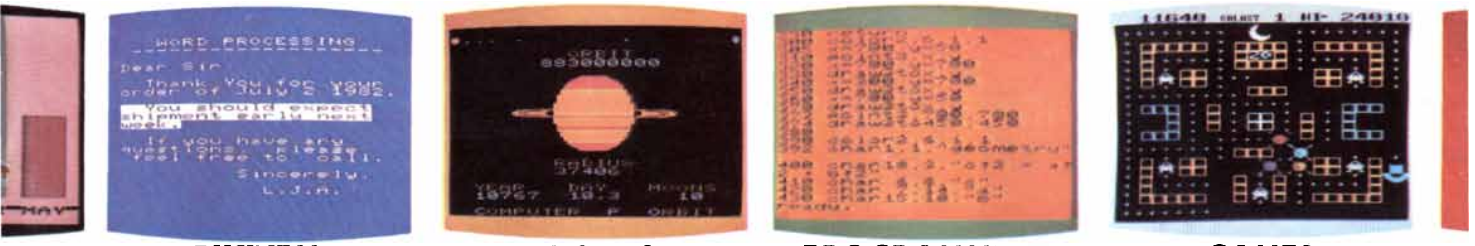
When a particle strikes a fixed target, much of the energy invested in accelerating the particle appears not as mass but as the kinetic energy of the product particles. If two particles collide head on, however, all their kinetic energy is available for conversion into mass. This is the principle of the storage ring.

In the simplest storage rings the colliding particles are electrons and their antiparticles, positrons. Because a particle and its antiparticle have the same mass but opposite charge, a single system of accelerator cavities and bending magnets can propel them around the ring in opposite directions. Where the counterrotating beams pass through each other, collisions are observed. The collision energy is the sum of the energies of the two beams. In the past 10 years or so some half-dozen electron-positron storage rings have been built. The largest of them has a collision energy of about 38 GeV and is being upgraded to 40 GeV, where particles incorporating a top quark might be accessible.

LEP, with electron and positron beams of 50 GeV each, will reach a collision energy of 100 GeV. Reaching that energy range is considered important because the mass of the Z^0 is predicted to be about 94 GeV. The Z^0 may be seen before LEP is completed in experiments now under way at CERN with a proton-antiproton storage ring. LEP, however, should give a much clearer view of the particle's properties. A second phase of the LEP project calls for the storage ring to be upgraded to a collision energy of 260 GeV. What will be seen at this energy is a matter of speculation, but it is entirely possible that new quarks and leptons will be discovered.

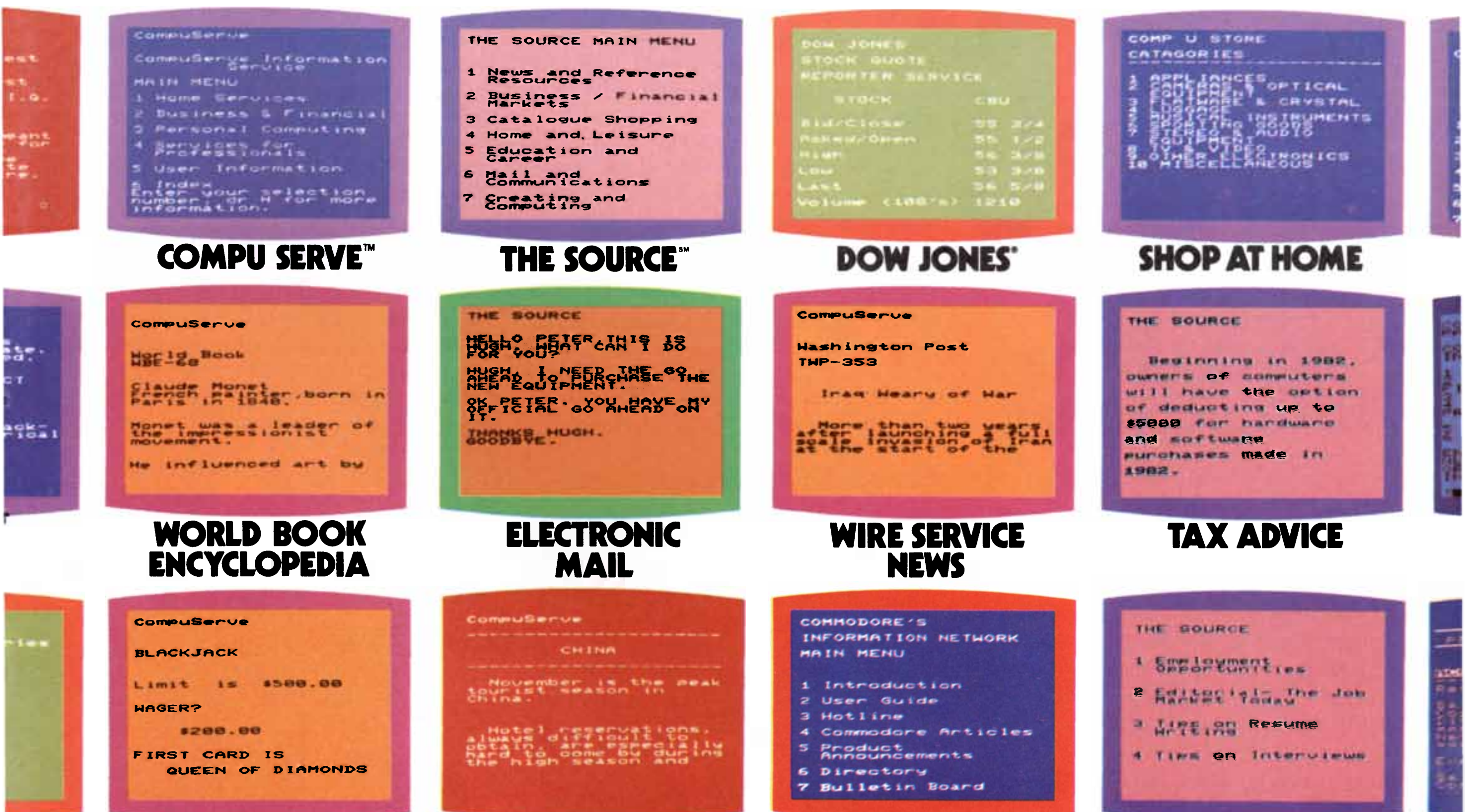
LEP will almost surely be the preeminent accelerator of the 1990's. Emilio Picasso of CERN, the director of the LEP program, has said that it will probably also be the last accelerator of its kind. The impediment to building a still more powerful synchrotron or storage ring is that charged particles following a curved trajectory emit the electromagnetic radiation called synchrotron radiation. The radiation dissipates energy that must be made up by supplying additional radio-frequency power. The energy loss increases as the fourth power of the beam energy. It can be reduced by building a ring with a larger circumference, but construction costs will rise.

IF YOU OWN A COMMODORE COMPUTER, YOU KNOW IT CAN DO ALL THIS.



BUSINESS — EDUCATION — PROGRAMMING — GAMES

BUT DID YOU KNOW FOR ABOUT $100, YOU CAN ALSO GET IT TO DO ALL THIS?



COMPU SERVE™ — THE SOURCE℠ — DOW JONES® — SHOP AT HOME

WORLD BOOK ENCYCLOPEDIA — ELECTRONIC MAIL — WIRE SERVICE NEWS — TAX ADVICE

GAMES — TRAVEL INFORMATION — COMMODORE INFO. NETWORK — EMPLOYMENT OPPORTUNITIES

The screens at the top of the page show a few examples of how versatile the VIC 20™ or Commodore 64™ can be with the addition of Commodore software.

The screens below them give you a few examples of how much more versatile they can be with the addition of a Commodore VICMODEM.

For around $100, the Commodore VICMODEM will turn your VIC 20 or Commodore 64 computer into a telecomputer.

To make matters even better, Commodore includes a few little extras (such as a free hour's time on the two most popular telecomputing services) that add up to a value of $197.50.* A nice return on an investment of about $100.

Most computer companies think it's reasonable to ask as much as $500 for a modem that'll give you telecomputing capabilities such as ours.

However, with a VICMODEM priced at around $100, we think we're being a lot more reasonable. Don't you agree?

commodore COMPUTER

U.S.A.–P.O. Box 500, Conshohocken, PA 19428; Canada– 3370 Pharmacy Avenue, Agincourt, Ontario, Canada M1W2K4. *Certain offers subject to change. CompuServe is a trademark of CompuServe, Inc. and H.&R. Block Co. Dow Jones News/Retrieval Service is a registered trademark of Dow Jones & Co., Inc. The Source is a service mark of Source Telecomputing Corporation, a subsidiary of Reader's Digest Corporation, Inc

Burton Richter of SLAC, who is directing the SLC project, argues that the cost of a circular collider is proportional to the square of the beam energy, whereas the cost of a linear collider is proportional to the energy itself. Therefore, he states, linear colliders will provide the cheapest way of reaching energies higher than the 260 GeV proposed for the second phase of LEP.

The SLC will be built by modifying the existing linear accelerator at SLAC, which is more than three kilometers long and already supplies electrons and positrons to two storage rings. A loop 600 meters in diameter is to be added at the end of the accelerator. Packets of electrons and positrons will be diverted into the loop, the electrons into one arc and the positrons into the other; they will collide on the far side. Electrons and positrons can be accelerated in the same direction by synchronizing their injection into the accelerator with the phase of the alternating radio-frequency fields in the cavities.

Since the SLC will provide a collision energy of about 100 GeV, it too could be used to examine the Z^0. It will be able to accommodate far fewer experiments than LEP, however, and there are no plans to upgrade it to higher energies. The main importance of the SLC may not be the physics experiments that could be done there. It is in itself an experiment in the design of a new type of particle accelerator, perhaps the first of a new generation.

Compatibility Genes

The histocompatibility antigens are molecules embedded in the cell's surface membrane that give an individual animal's tissue its own chemical identity. When a patch of skin or an organ is transplanted, it is these antigens that are recognized as nonself by the host's immune system, which proceeds to attack the foreign tissue. The rejection is most vigorous when the donor of tissue and the recipient have different versions of certain genes: those of the major histocompatibility complex, which in man occupies a short segment of chromosome 6 and is called the *HLA* region. In the mouse similar genes occupy the *H*-2 region of chromosome 17. The histocompatibility genes cannot have been conserved in evolution expressly to foil transplantation, and indeed they have been found to govern a wide range of biological functions. Among other things, it now appears, they determine characteristic body odors by which animals recognize one another.

In 1974 some staff members of the mouse-breeding facility at the Memorial Sloan-Kettering Cancer Center noticed that the social behavior of mice seemed to be influenced by their *H*-2 type. In a number of cages an inbred male mouse was kept with two females. One female was of the same inbred strain as the male and was therefore syngeneic, or genetically identical. The other female was *H*-2 congenic: genetically identical with the male except for the region of chromosome 17 carrying the *H*-2 genes. The male and the female whose *H*-2 genes were different tended to consort and nest with each other, excluding the syngeneic female. As it happened, Lewis Thomas, who was then president of Memorial Sloan-Kettering, had recently speculated that histocompatibility genes might give individual animals a characteristic scent. The combined stimulus of the breeding-trio observation and Thomas' suggestion led investigators at the center to join with a group at the Monell Chemical Senses Center in Philadelphia to learn whether mating preferences in mice are influenced by *H*-2 type, with recognition mediated by scent.

Their first experiment was a controlled test of the original breeding-trio observation. A male was caged with a syngeneic female and an *H*-2 congenic one. Thousands of trials confirmed that there is an overall bias in favor of mating between mice having different *H*-2 genes. Such pairing leads to *H*-2 heterozygosity in the offspring (that is, the offspring bear different sets of *H*-2 genes on their two copies of chromosome 17), presumably a beneficial trait.

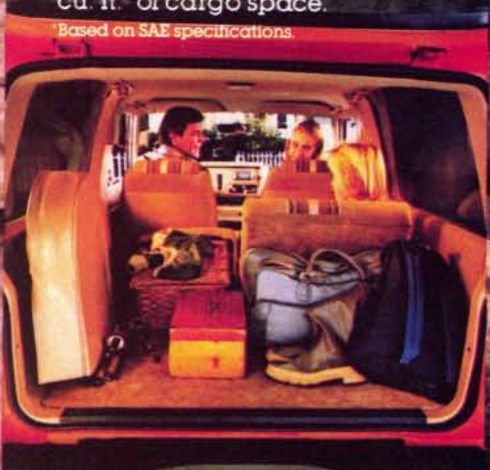
The next step was to establish the mode of *H*-2 recognition, which seemed likely to be olfactory but could have been based on taste, or even on hearing since mice are known to communicate by ultrasonic signals. The ability of mice to distinguish between two *H*-2 types was tested in a *Y*-shaped maze. The test mouse was placed in the stem of the *Y*. Each arm of the *Y* was scented by air flowing through a chamber housing either a male mouse or its urine. The two odor-source mice were *H*-2 congenic with each other. Each thirsty test mouse was rewarded with a drop of water when it entered the arm of the maze scented by a mouse having a particular *H*-2 type (or by the urine of such a mouse). The experiment was strictly controlled to eliminate unknown adventitious signals and unintentional prompting by the operators of the maze. All the test animals, regardless of their genetic makeup, learned to distinguish between the *H*-2 types presented to them.

The *Y*-maze experiments have led to a number of conclusions, which were reported at a meeting of the International Society for Oncodevelopmental Biology and Medicine by Edward A. Boyse, Gary K. Beauchamp, Kunio Yamazaki, Judith Bard and Thomas. The sensory recognition of *H*-2 types is confirmed. Particularly in view of the successful training to distinguish urine alone, a prime mode of recognition (not neces-

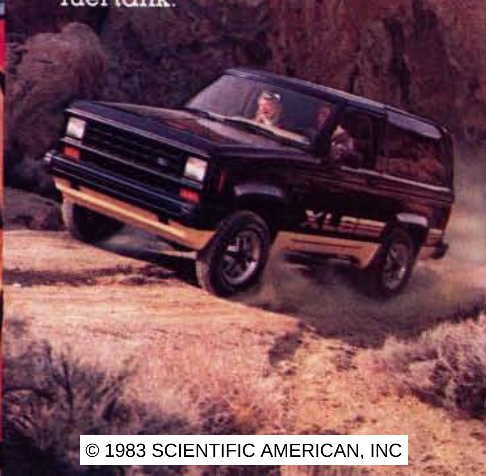

Since 1715.
One of the world's more civilized pleasures.
J&F MARTELL
VS
MARTELL
FONDÉE EN 1715
J & F. Martell
COGNAC
750 ML
(25.4 FL.OZ)
80 PROOF
MARTELL & Cº COGNAC
PRODUCED AND BOTTLED IN FRANCE
To send a gift of Martell in the U.S., call 800-528-6148.
Imported by The Jos. Garneau Co., N.Y., N.Y. © 1981.

Trim new Ford 4-wheeler seats four in comfort. Split fold-down rear seats add extra versatility, plus up to 64 cu. ft.* of cargo space.

*Based on SAE specifications.



V-6 engine standard, a V-6 with power the S-10 Blazer can't match. Extra pull for mud and snow. Big 23-gal. fuel tank.



Flip to 4-wheel drive from the driver's seat with auto-lock hub option. In snow or bad weather, shift to 4WD traction and relax.


BRONCO II
IT'S A BRAND NEW KICK!!
FORD
FORD DIVISION
Ford
Built Ford tough, with husky frame and chassis, tested Twin-Traction Beam front suspension. Proven manual or optional auto-lock hubs.
New sporty trim size makes driving and parking easy. Roomy, comfortable interior. Plus independent front suspension for controlled ride.
Perfect size for 4-wheeling– maneuverable Bronco II turns in less space than the S-10 Blazer or Jeep CJ's. Try it . . . for a brand-new kick!
Get it together–Buckle up.

sarily the only one) is olfaction. The recognition is not dependent on sex. Although mice choosing a mate seem generally to prefer animals of a different *H*-2 type, in the maze there is no obvious bias of untrained mice toward either a similar or a different type.

The workers at Monell and Memorial Sloan-Kettering are continuing their study, which opens many lines of inquiry. What other genes might define body odors? Already there is evidence that variation in the genetically allied region called *Qa:Tla* adjacent to the *H*-2 genes labels the urine of mice distinctively. What effects might the *H*-2 region have on aspects of reproductive behavior other than mating preference? There is evidence that it affects the phenomenon known as pregnancy block. The implantation of a fertilized ovum may be prevented when a female is exposed to a strange male soon after mating; there is now evidence that pregnancy block is more likely when the strange male is *H*-2-congenic, rather than syngeneic, with the stud male. Is there a relation between olfactory recognition and immunologic recognition? The major histocompatibility genes seem to be involved in both systems; olfaction, like the immune response, depends on cell-surface receptors, but the nature of the olfactory receptors and of the signals they transmit to the brain is still far from being understood.

Hybrid for Hybrid

A recent agricultural agreement shows that the transfer of technology between richer countries and poorer ones can go in both directions. A hybrid variety of rice developed in China and grown widely there is now under study in the U.S. In exchange a hybrid variety of cotton developed in the U.S. is being raised experimentally in China. The swap was arranged in 1980 by Armand Hammer of the Occidental Petroleum Corporation. The U.S. study is being done by Ring Around Products, Inc., an agricultural subsidiary of Occidental and the developer of the hybrid cotton.

The rice, a medium-grain glutinous variety known as Xian three-line, has given its American growers yields that are from 20 to 50 percent higher than those of commercial nonhybrid medium-grain varieties. In China, where six million hectares are now being planted exclusively with Xian three-line, it has reportedly increased overall rice production by some 13 million metric tons in the past five years. The plants are drought-resistant and suitable for a wide range of climates. According to the Chinese Science and Technology Commission, the grain has a protein content of from 9 to 11 percent.

The main impediment to commercial exploitation of the new rice is the cost of seed production, which is higher than it is for other hybrid crops such as corn. Growing seed for hybrid rice is a labor-intensive activity calling for considerable manipulation of the individual plants. Workers at Ring Around Products are looking for ways to mechanize seed production. A second objective is to further hybridize the variety so that it will bear nonglutinous long grains, the kind of rice preferred by American and European consumers, while preserving the Chinese hybrid's large root system and consequent resistance to drought.

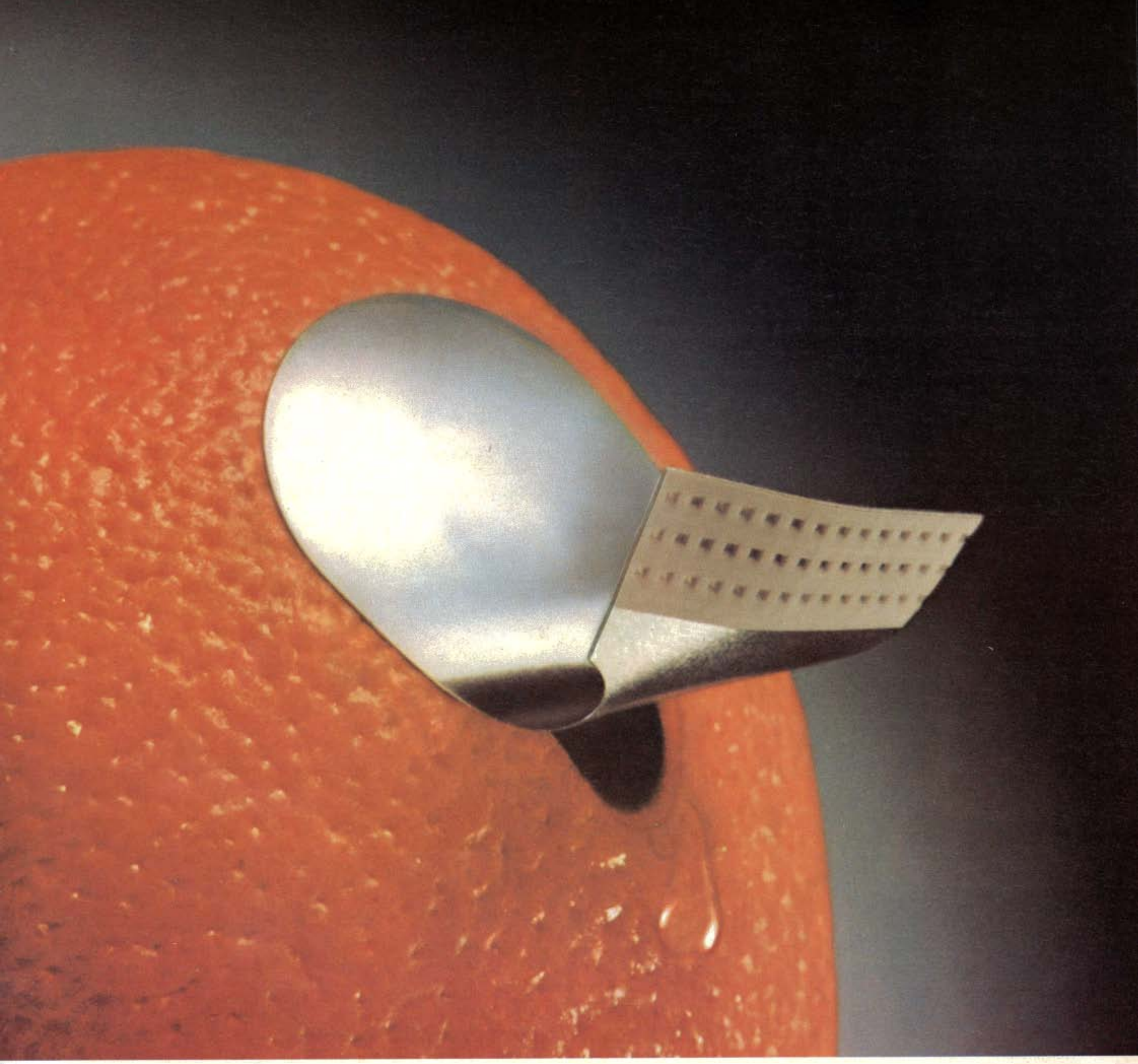
Chilling Sound

Cycles of compression and rarefaction are the basis both of sound waves and of refrigeration machinery. Can a refrigerator be run by the sound from a loudspeaker? An acoustic heat engine devised by John C. Wheatley of the Los Alamos National Laboratory has demonstrated the principle.

If a piston is set in reciprocating motion in one end of a tube that is closed at the other end, heat builds up in the tube. Wheatley mounts a stack of thermally insulated fiberglass plates in such a tube and couples a loudspeaker to the piston. The loudspeaker emits sound at acoustic frequencies (from 150 to 1,000 hertz). Soon the end of the stack nearest the piston begins to cool while the other end heats up. Within a few minutes the difference in temperature can become quite large, perhaps 100 degrees Celsius.

Wheatley's experimental tube is a meter long and three centimeters in diameter. The plates are 10 centimeters long and are separated from one another by one millimeter. The tube is filled with a gas (in Wheatley's experiments it has been helium) and then closed with the piston. As the gas oscillates along the surface of each plate in response to the action of the piston, heat flows out of one end of the plate, through the gas along the surface of the plate and back into the other end. As it comes out of the cold end of the plate it is moving diffusively, that is, according to the laws of thermal conduction. As it moves along the plate it is flowing hydrodynamically, that is, because the gas is moving. It flows diffusively again as it passes back into the other end of the plate.

According to Wheatley, the acoustic heat engine can utilize a variety of thermally active materials. The thermal cycle depends on the configuration of the device and on the fact that thermal contact between the parts is imperfect. On making a practical machine, Wheatley says: "To get a large temperature difference you must reduce the heat input to the cold end. You can reduce external heat as much as you want, but a lot of heat is generated in the tube itself. We haven't tried yet to build something useful, but we're thinking about it."



WHERE DO BLACK DOCTORS COME FROM?

Nearly half of all black doctors practicing in the U.S. today are graduates of Meharry.

Sure, there are black medical students at Harvard and Stanford and Johns Hopkins. But no single private medical school in the U.S. produces as many black physicians as Meharry Medical College.

Ninety percent of our students require financial aid to some degree. To make more funds available for scholarships and student loans, and to maintain the quality of our training, our goal is to raise $27 million.

Meharry must depend largely on private and corporate contributions. We are eligible for only limited funding under Federal and State programs—and, as a professional school, receive no resources from the United Negro College Fund.

Won't you help? Because without Meharry, a lot of qualified youngsters couldn't even consider a medical career.

OPEN WIDE FOR
MEHARRY
MEDICAL COLLEGE

Send your tax-deductible contribution, in any amount, to **Meharry Medical College, Second Century Fund, Nashville, TN 37208.**
The need is urgent, and your support is appreciated.

First!

Asia's first international airline would like to invite you to experience some important "firsts" and a touch of class you won't find anywhere else.

It's our First Class. And with award-winning Nouvelle Cuisine, more full-length beds, and daily flights from the West Coast to Manila, it's no wonder Asia's first airline still is.

Class!

Philippine Airlines

Asia's first airline still is.

Amsterdam Athens Bahrain Bangkok Brisbane Canton Dhahran Dubai Frankfurt Hong Kong Honolulu Jakarta Karachi Kuala Lumpur Kuwait London Los Angeles Manila Melbourne Peking Port Moresby Rome San Francisco Singapore Sydney Taipei Tokyo

Mitochondrial DNA

The cell's energy-generating organelle has its own genetic system. The organization of its genes and even its version of the "universal" genetic code are different from those of the nucleus and bacteria

by Leslie A. Grivell

Some one and a half or two billion years ago, when the earth was still poor in oxygen, a primitive bacterium that made a precarious living from the anaerobic fermentation of organic molecules engulfed a smaller cell that had somehow evolved the ability to respire. The event was a turning point in organic evolution. Respiration liberates far more energy than fermentation, and the growing abundance of oxygen in the atmosphere must have been the driving force behind a symbiotic relation that developed between the two cells, with the aerobic cell generating energy in return for shelter and nutrients from its larger host.

In time the engulfed cell and others like it were to become subcellular organelles, passed on by host cells to their progeny. Eventually the host cells themselves changed, developing other subcellular structures and internal membranes and segregating their genetic material in chromosomes within a nucleus. These cells were the ancestors of all modern eukaryotic (nucleated) cells: protozoans and algae and the individual cells of fungi, plants and animals. The present-day descendants of those ancient symbiotic respiring bacteria are the mitochondria, the power plants of the eukaryotic cell.

Mitochondria are oval or worm-shaped organelles, about half a micrometer in diameter and from two to five micrometers long; this is roughly the size of many modern bacteria. The mitochondrion has an outer membrane and an extensively folded inner membrane that encloses a fluid matrix. The organelle is the site of oxidative phosphorylation, the primary source of cellular energy. In the fluid matrix organic molecules derived from the breakdown of foodstuffs are oxidized in a series of chemical reactions known as the citric acid cycle. Electrons removed in the course of oxidation are passed along a chain of respiratory-enzyme complexes arrayed in the inner membrane, driving the phosphorylation of adenosine diphosphate to form adenosine triphosphate (ATP), the universal energy carrier of cells. The cytoplasm of eukaryotic cells (the region outside the nucleus) contains from a few mitochondria to many hundreds; the greater the energy demands placed on the cell are, the more mitochondria it has.

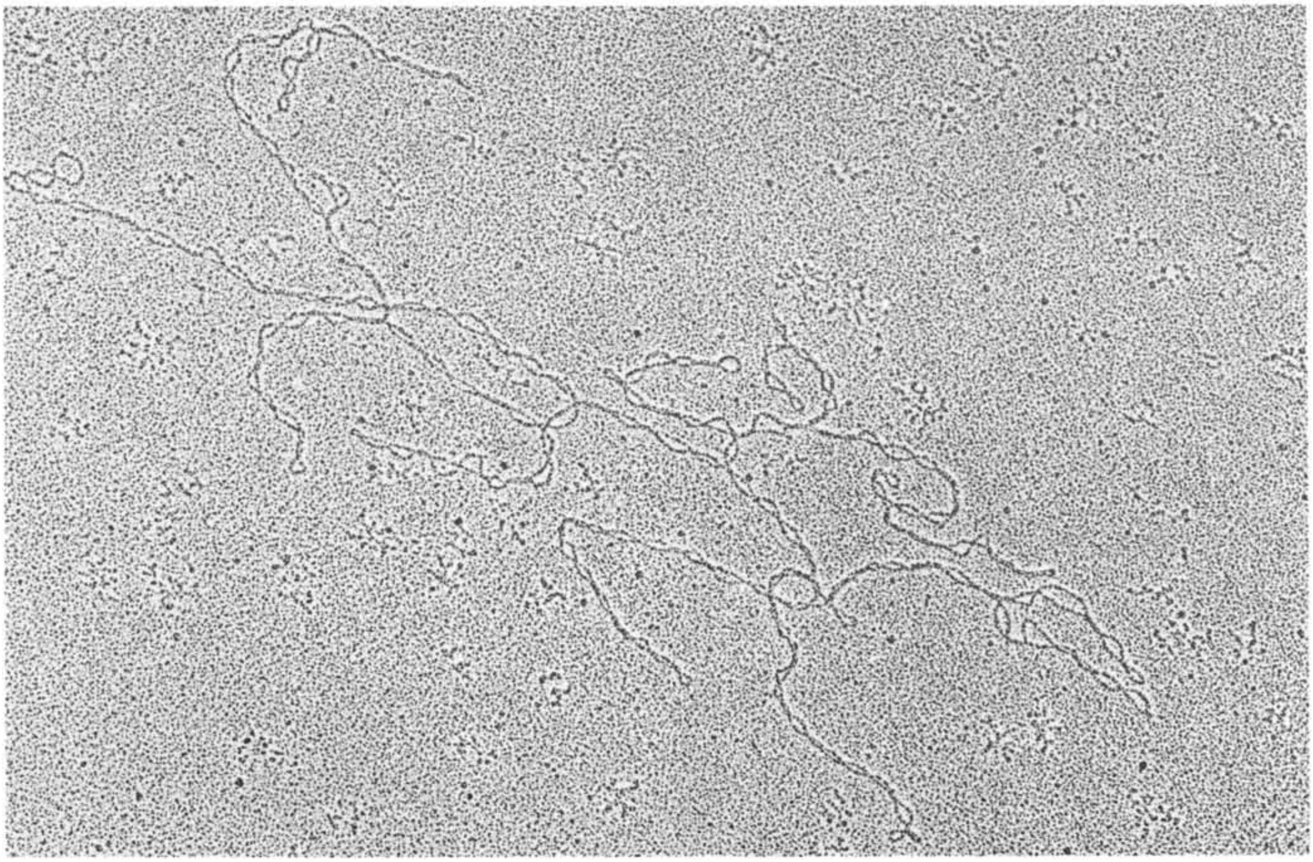
Perhaps because they are descendants of a free-living bacterium, mitochondria have their own genetic material and the machinery to express it. The possibility that the organelles might have their own genes, distinct from the genome in the cell nucleus, was raised as early as 1949, when Boris Ephrussi found that the ability of baker's yeast to carry out oxidative phosphorylation seemed to be controlled by some factor in the cytoplasm rather than in the nucleus. It was only in 1966, however, that the first vertebrate mitochondrial DNA (that of the chick) was isolated and characterized. Since then the mitochondrial DNA's of many organisms have been under intensive study in a number of laboratories in Europe and the U.S. The intriguing questions have been: What does mitochondrial DNA do? How is it organized and expressed? Why should the cell bother to maintain a second genetic system alongside the one in the nucleus? Answers to the first two questions are emerging, and they make it possible at least to speculate about the broader third question.

Important new data have been reported in the past two years, thanks largely to the application of molecular genetics and powerful new techniques for rapidly determining the sequence of nucleotides (the building blocks of DNA) in a genome. The complete nucleotide sequence of human mitochondrial DNA was published early in 1981 by B. G. Barrell and his colleagues in Frederick Sanger's laboratory at the Medical Research Council Laboratory of Molecular Biology in Cambridge; at the same time Giuseppe Attardi of the California Institute of Technology analyzed the transcription of human mitochondrial DNA into RNA. Soon the sequence of bovine mitochondrial DNA was published by Barrell's group and that of the mouse by David A. Clayton and his colleagues at Cal Tech. Meanwhile the sequence of a large part of the yeast mitochondrial genome has been unraveled by Alexander Tzagoloff of Columbia University and his group. What is becoming clear is that the diverse mitochondrial DNA's are unlike any other known DNA in several fundamental ways. Their unusual features may be the last traces of the primitive aerobic organism that long ago colonized fermenting bacteria and thereby initiated the eukaryotic revolution.

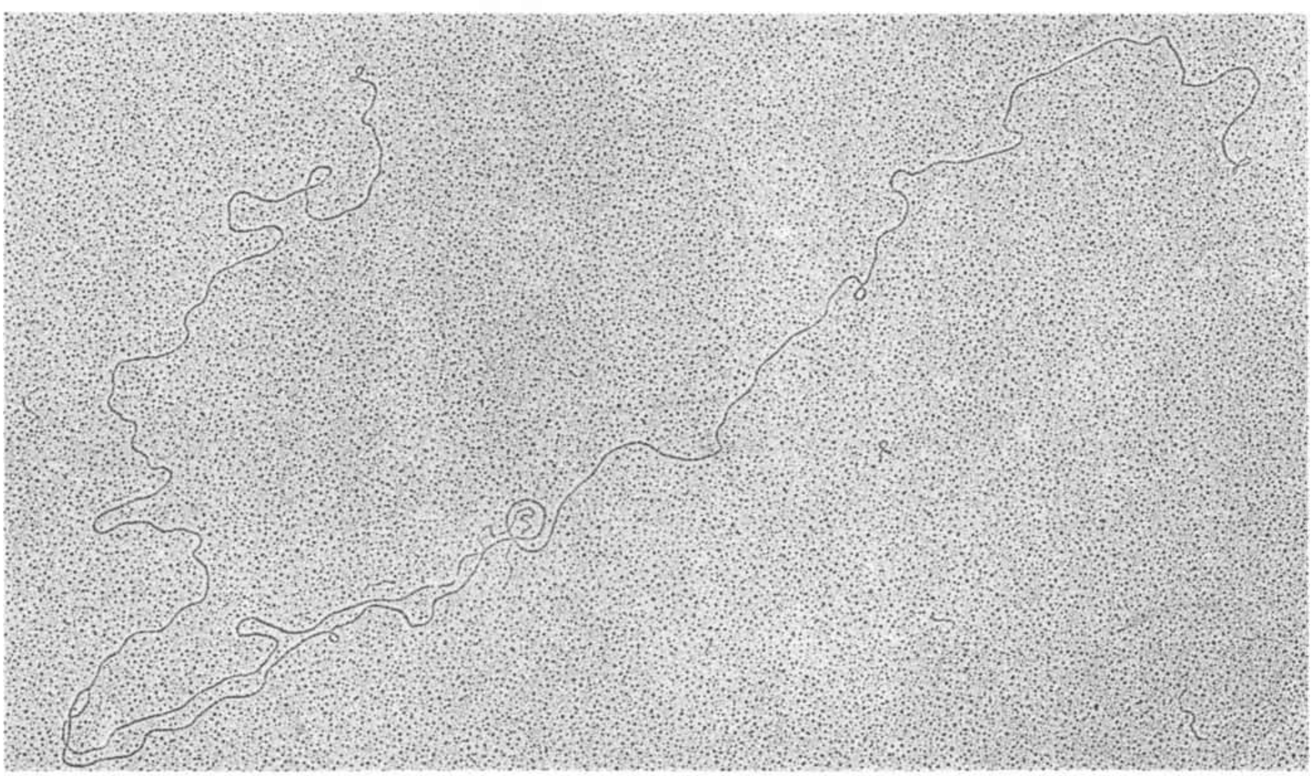
The Genome and Its Proteins

Mitochondrial DNA is like that of bacteria in that it consists of a single double helix of naked DNA, whereas the nuclear DNA of eukaryotes is tightly complexed with protein and is apportioned among a number of discrete chromosomes. Except in some protozoans the DNA helix is formed (again as in bacteria) into a circle. In most species, however, the mitochondrial genome is much smaller than that of the simplest bacteria; it is closer in size to the genome of many viruses. The size varies widely with the organism. Most animal mitochondria have only some 15,000 nucleotides, but in yeast the mitochondrial DNA is five times as long and in plants it is as much as five times longer still. Much of the difference in length between animal and yeast DNA's can be accounted for by the great difference in the way the same basic set of genes is organized. Whether this holds true for the mitochondrial DNA of plants is as yet not known.

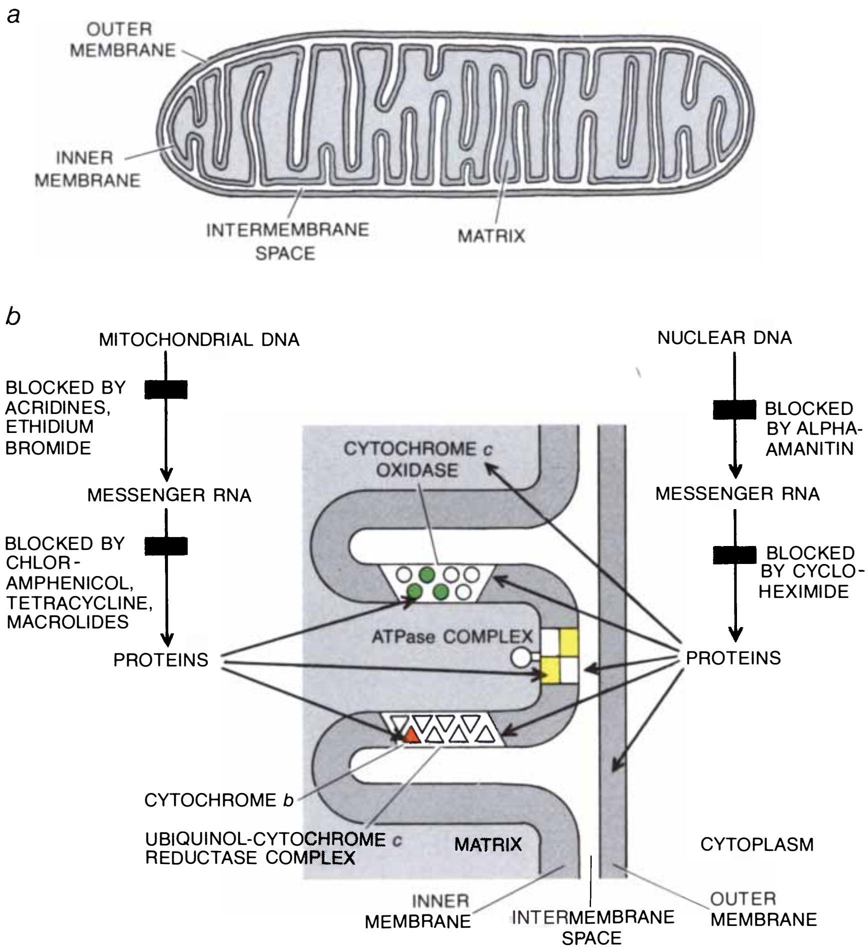
Although the mitochondrion has its own genetic material, it is not genetically self-sufficient: most of the proteins that account for the structure and function of the organelle are encoded in the nucleus. The mitochondrion's own DNA specifies the RNA of its ribosomes (the jigs on which proteins are assembled) and its transfer RNA's (which carry specific amino acids to the



MITOCHONDRIAL DNA from the brewer's yeast *Saccharomyces carlsbergensis* is enlarged some 60,000 diameters in an electron micrograph made by Ernst F. J. Van Bruggen of the State University at Groningen in the Netherlands. The molecule is a "supercoiled" circle of double-strand DNA with a circumference of 26 micrometers. It is made up of some 75,000 nucleotides (the building blocks of DNA).



MITOCHONDRIAL GENOME of the protozoan *Tetrahymena pyriformis* is a linear molecule rather than a circle. The DNA is enlarged some 26,000 diameters in this micrograph made by Annika C. Arnberg and Van Bruggen. It is seen at an early stage in replication. A replication "eye" has formed near the middle of the molecule and is being extended as new DNA is synthesized along each of the strands.



STRUCTURE AND FUNCTION of the mitochondrion are defined jointly by genes in the cell nucleus and in the mitochondrion itself. The organelle, seen in longitudinal section (*a*), has an outer membrane and a deeply folded inner membrane that encloses a fluid matrix. In all organisms examined so far the DNA of the mitochondrion encodes the organelle's own ribosomal RNA's and transfer RNA's, which are elements of the protein-synthesizing machinery. Other mitochondrial genes code for proteins. In yeast just six proteins are known to be specified by the mitochondrial genome (*b*): three subunits of the enzyme cytochrome *c* oxidase (*green*), two subunits of the enzyme ATPase (*yellow*) and the protein cytochrome *b* (*red*). All six take part in oxidative phosphorylation, the main source of cellular energy. Most proteins constituting the organelle or needed to operate its genetic system are encoded in the nucleus. It is possible to learn whether the nuclear genetic system or the mitochondrial one is responsible for a particular synthesis by blocking one of the transcription or translation systems with a selective drug.

ribosomes for assembly into proteins). As for proteins, so far mitochondrial DNA's have been found to make messenger RNA that codes only for a few subunits of the respiratory-enzyme complexes in the inner membrane of the organelle. The rest of the inner-membrane proteins as well as those of the matrix (including the enzymes of the citric acid cycle) and those of the outer membrane are all encoded by nuclear genes; they are synthesized by ribosomes in the cytoplasm and imported into the mitochondrion. In addition to supplying the mitochondrion with most of its structural materials the nucleus encodes all the enzymes needed to replicate DNA and to transcribe it into RNA; the nucleus also encodes the proteins that combine with RNA to form the ribosomes and specifies other factors needed for the synthesis of mitochondrial proteins. In other words, the mitochondrial genetic system is maintained only through a considerable investment on the part of the nucleus and the cell's own protein-synthesizing machinery.

In the living cell the disparate contributions of the two genetic systems are coordinated, but there is some play in the coupling; the cell can live for a time while the genetic activity of either the nucleus or the mitochondria is blocked. Mitochondrial protein synthesis, for example, is blocked by antibacterial antibiotics such as chloramphenicol, tetracycline and the macrolide group; it is not affected by cycloheximide, which inhibits the activity of cytoplasmic ribosomes. One can arrange things so that proteins synthesized while one system is blocked are labeled with a radioactive isotope, and thus determine which system does what.

Labeling experiments have identified the products of mitochondrial DNA in a number of organisms. To identify the genes that encode the products and to map the positions of the genes on the circle of DNA was a harder task, however, or at least it was until the new methods for sequencing DNA were developed. The identification and mapping of genes by genetic analysis depends on noting and tracking mutations that interfere with a gene's activity without killing the organism. In the absence of oxygen a yeast can get by on fermentation alone, and so it can survive without functional mitochondria. It was therefore possible to note and characterize mutations in yeast that are lethal in other organisms, and as a result the mitochondrial DNA of yeast was the first to come under intensive study.

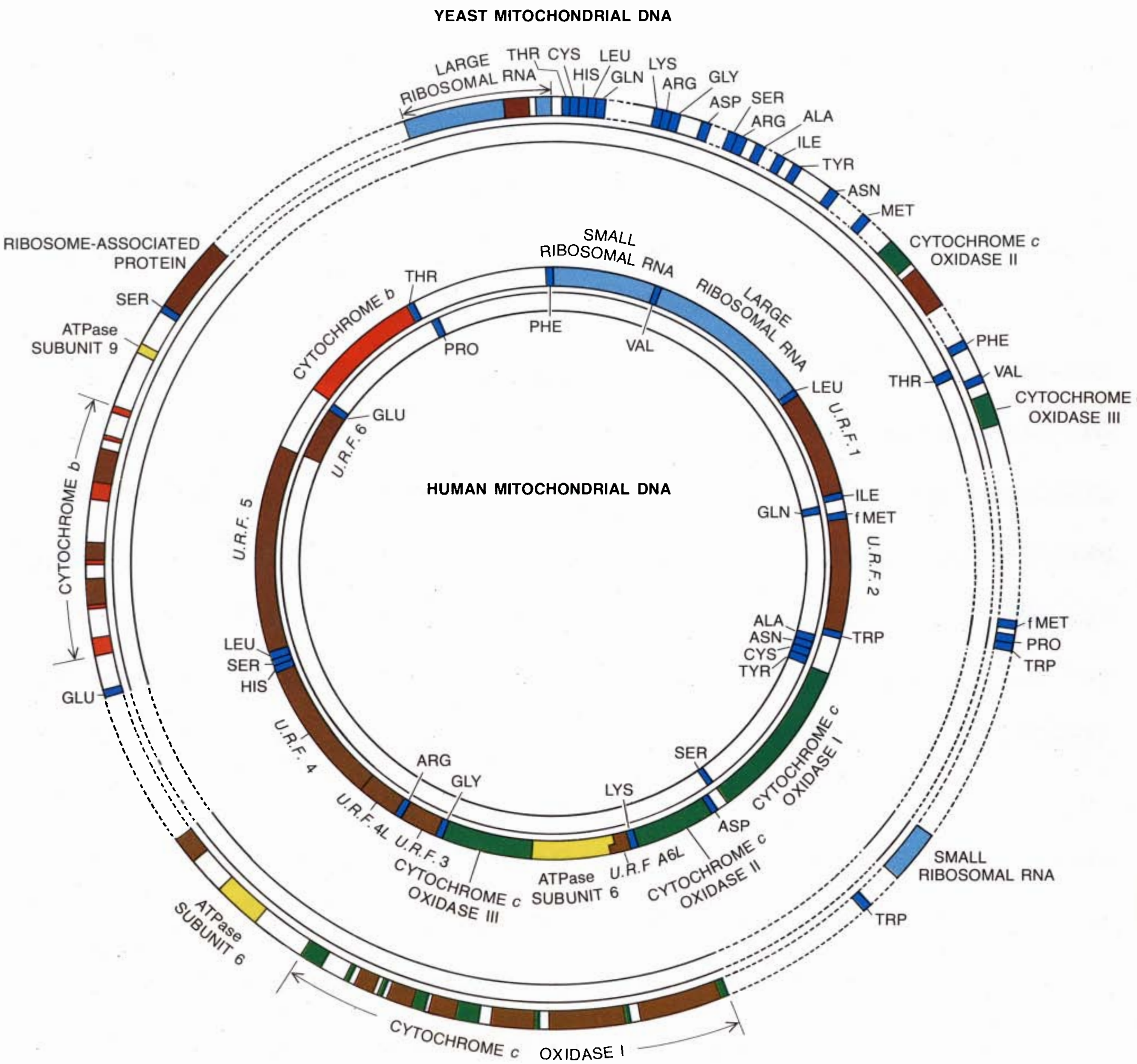
By now a combination of genetic analysis and DNA sequencing has provided inventories of the genes of both the human and the yeast mitochondrial genomes, together with detailed maps of their arrangement. Mitochondria from yeast and from man have essentially the same genes, but the organization of those genes is strikingly different.

Genome Organization

Each nucleotide of DNA is characterized by one of four bases, and the nucleotide sequences of the two strands of the double helix are complementary: adenine (*A*) on one strand pairs with thymine (*T*) on the other strand, and guanine (*G*) pairs with cytosine (*C*). Triplets of bases on one strand (the coding strand) serve as code words (codons) that, having been transcribed into RNA, specify either a particular amino acid or a signal for the initiation or termination of translation. In most eukaryotic chromosomal DNA the codons are a minority. The strings of triplets constituting a gene are flanked by long noncoding spacers and signaling regions that are not transcribed into RNA and by other stretches that are transcribed into messenger RNA but are not translated; the genes themselves are often split into pieces by noncoding intervening sequences ("introns").

Human mitochondrial DNA, in contrast, is a model of economy. It has only 16,569 base pairs, almost every one of which seems to have been put to use, with one strand or the other serving in turn as the coding strand. Indeed, the genes are so tightly packed that there are few if any noncoding nucleotides between them. There are genes for two ribosomal RNA's, for 22 transfer RNA's and for five membrane proteins; in addition there are eight unassigned reading frames. (These are regions where successive triplets of nucleotides can be read as codons specifying amino acids and where the sequence of those triplets is not interrupted by "stop" codons. Such regions must specify a protein, but their protein product has not yet been identified.) None of the genes contain introns. Very little of the DNA codes for messenger RNA that is not subsequently translated into protein.

In both bacterial genomes and nonmitochondrial eukaryotic genomes the transcription of each gene is individually controlled by an interaction between transcribing enzymes and signaling sequences in noncoding regions of the DNA. The scarcity of such DNA in the human mitochondrial genome suggests transcription must be regulated in some unusual way. How this is managed is beginning to come clear as the result of painstaking studies by Attardi and his colleagues at Cal Tech. It appears there is only one major promoter, or initiation site for RNA synthesis, on each strand of human mitochondrial DNA. The primary product of transcription is therefore a full-length RNA copy of each strand. The two primary transcripts are then cleaved to yield the ribosomal RNA's, the transfer RNA's and the messenger RNA's that are subsequently translated into protein; the cleavage takes place even as the RNA is being synthesized. In the absence of noncoding DNA, what sites in the pri-



HUMAN AND YEAST MITOCHONDRIA have essentially the same set of genes, but their organization is different. The two DNA's are shown here as concentric double circles (one circle for each strand of the double helix) but at different scales: the yeast DNA molecule is actually about five times as long as the human one. The full nucleotide sequence of the human genome has been determined by B. G. Barrell and his colleagues; most of the sequence for the yeast *Saccharomyces cerevisiae* is also known, but some stretches remain unsequenced (*broken lines*). Colored segments show the extent of genes for known proteins (*green, yellow and red*), for ribosomal RNA's (*light blue*) and for transfer RNA's (*dark blue*), each of which is labeled with the abbreviation for its specific amino acid; the brown segments indicate unassigned reading frames (*U.R.F.*), which are presumably genes for proteins that have not been identified. The human genome is extremely compact, with little noncoding DNA between genes. In the yeast genome, on the other hand, there are long noncoding stretches. Several genes are interrupted by intervening sequences ("introns"), some of which in turn include unassigned reading frames.

→ CYTOCHROME *c* OXIDASE I →

TCATTCACTGATTTCCCCTATTCTCAGGCTACACCCTAGACCAAACCTACGCCAAAATCCATTTCACTATCATATTCATCGGCGTAAATC

7,100

TAACTTTCTTCCCACAACACTTTCTCGGCCTATCCGGAATGCCCCGACGTTACTCGGACTACCCCGATGCATACACCACATGAAACAT

7,200

CCTATCATCTGTAGGCTCATTCATTTCTCTAACAGCAGTAATATTAATAATTTTCATGATTTGAGAAGCCTTCGCTTCGAAGCGAAAAGT

7,300

CCTAATAGTAGAAGAACCCTCCATAAACCTGGAGTGACTATATGGATGCCCCCCACCCTACCACACATTCGAAGAACCCGTATACATA

7,400 ← TRANSFER RNA (SER) ← →

AAATCTAGACAAAAAAAGGAAGGAATCGAACCCCCCAAAGCTGGTTTCAAGCCAACCCCATGGCCTCCATGACTTTTTCAAAAAGGTA

→ TRANSFER RNA (ASP) → CYTOCHROME *c* OXIDASE II →

TTAGAAAAACCATTTCATAACTTTGTCAAAGTTAAATTATAGGCTAAATCCTATATATCTTAATGGCACATGCAGCGCAAGTAGGTCTA

7,600

CAAGACGCTACTTCCCCTATCATAGAAGAGCTTATCACCTTTCATGATCACGCCCTCATAATCATTTTCCTTATCTGCTTCCTAGTCCTG

7,700

TATGCCCTTTTCCTAACACTCACAACAAAACTAACTAATACTAACATCTCAGACGCTCAGGAAATAGAAACCGTCTGAACTATCCTGCC

CGCCATCATCCTAGTCCTCATCGCCCTCCCATCCCTACGCATCCTTTACATAACAGACGAGGTCAACGATCCCTCCCTTACCATCAAA

7,800

TCAATTGGCCACCAATGGTACTGAACCTACGAGTACACCGACTACGGCGGACTAATCTTCAACTCCTACATACTTCCCCCATTATTCCT

7,900

AGAACCAGCCGACCTGCGACTCCTTGACGTTGACAATCGAGTAGTACTCCCGATTGAAGCCCCCATTCGTATAATAATTACATCACAA

8,000

GACGTCTTGCACTCATGAGCTGTCCCCACATTAGGCTTAAAAACAGATGCAATTCCCGGACGTCTAAACCAAACCACTTTCACCGCTA

8,100

CACGACCGGGGGTATACTACGGTCAATGCTCTGAAATCTGTGGAGCAAACCACAGTTTCATGCCCATCGTCCTAGAATTAATTCCCCT

8,200 TRANSFER RNA (LYS) →

AAAAATCTTTGAAATAGGGCCCGTATTTACCCTATAGCACCCCCTCTACCCCCTCTAGAGCCCACTGTAAAGCTAACTTAGCATTAACC

U.R.F. A6L → 8,300

TTTTAAGTTAAAGATTAAGAGAACCAACACCTCTTTACAGTGAAATGCCCCAACTAAATACTACCGTATGGCCCACCATAATTACCCCC

8,400

ATACTCCTTACACTATTCCTCATCACCCAACTAAAAATATTAAACACAAACTACCACCTACCTCCCTCACCAAAGCCCATAAAAATAAA

ATPase SUBUNIT 6 →

AAATTATAACAAACCCTGAGAACCAAAATGAACGAAAATCTGTTCGCTTCATTCATTGCCCCCACAATCCTAGGCCTACCCGCCGCAG

8,500

TACTGATCATTCTATTTCCCCCTCTATTGATCCCCACCTCCAAATATCTCATCAACAACCGACTAATCACCACCCAACAATGACTAATCA

8,600

AACTAACCTCAAAACAAATGATAACCATACACAACACTAAAGGACGAACCTGATCTCTTATACTAGTATCCTTAATCATTTTTATTGCC

8,700

ACAACTAACCTCCTCGGACTCCTGCCTCACTCATTTACACCAACCACCCAACTATCTATAAACCTAGCCATGGCCATCCCCTTATGAGC

8,800

GGGCACAGTGATTATAGGCTTTCGCTCTAAGATTAAAAATGCCCTAGCCCACTTCTTACCACAAGGCACACCTACACCCCTTATCCCC

8,900

ATACTAGTTATTATCGAAACCATCAGCCTACTCATTCAACCAATAGCCCTGGCCGTACGCCTAACCGCTAACATTACTGCAGGCCACC

9,000

TACTCATGCACCTAATTGGAAGCGCCACCCTAGCAATATCAACCATTAACCTTCCCTCTACACTTATCATCTTCACAATTCTAATTCTAC

9,100 →

TGACTATCCTAGAAATCGCTGTCGCCTTAATCCAAGCCTACGTTTTCACACTTCTAGTAAGCCTCTACCTGCACGACAACACATAATGA

→ CYTOCHROME *c* OXIDASE III → 9,200

CCCACCAATCACATGCCTATCATATAGTAAAACCCAGCCCATGACCCCTAACAGGGGCCCTCTCAGCCCTCCTAATGACCTCCGGCC mary transcripts are recognized by the cleaving enzymes? Transfer-RNA transcripts flank just about every longer gene, and cleavage takes place precisely at the beginning and end of each transfer-RNA sequence. This suggests that the primary transcript folds in such a way that the distinctive shape of the transfer RNA's provides the recognition sites for the RNA-processing enzymes.

The lack of signaling space around the messenger RNA's similarly calls for unusual strategies to control translation. The signal for the initiation of translation is the same as the codon for the amino acid methionine. Nonmitochondrial messenger RNA's are preceded by sequences that serve as recognition sites for ribosomes, which bind there and thus label the following methionine codon as a "start" signal. Human mitochondrial ribosomes must be able somehow to recognize certain methionine codons as start codons without benefit of a preceding recognition site.

The problem of how to stop protein synthesis at the end of the gene is more complex. In nonmitochondrial messenger RNA's the termination of translation is signaled by one of several "stop" codons, among them *UAA* (*U* stands for uracil, the RNA analogue of thymine). Most of the human mitochondrial messengers either lack a stop codon (because they run into and overlap a following gene, usually one for a transfer RNA) or they have a stop codon that is truncated by the cleavage (mediated by transfer RNA's) described above. Those genes all end with either *U* or *UA*. How then is the stop codon supplied? The human mitochondrial messenger RNA, like most other eukaryotic messengers, is "polyadenylated," or given a long tail of adenine nucleotides, immediately after cleavage. The poly-*A* tail provides just what the translation machinery requires: a *UAA* stop signal.

Yeast mitochondrial DNA is as extravagant as its human counterpart is economical. The yeast genome has not yet been completely sequenced, but the DNA appears to code for two ribosomal RNA's, some 25 transfer RNA's, six membrane proteins and more than 10 unidentified proteins. In yeast (specifically in *Saccharomyces cerevisiae,* or baker's yeast) the genes are scattered over five times as much DNA. Coding sequences are separated by vast stretches of DNA consisting almost entirely of the bases *A* and *T.* At least three of the genes are split by introns. Long sequences at the beginning and end of genes are transcribed but not translated.

The Code Is Not Universal

Each codon in messenger RNA is "read" by the anticodon of a transfer RNA that brings a specific amino acid to take its place in a protein chain. The genetic code relating each codon of DNA (or its messenger-RNA transcript) to a particular amino acid is exactly the same in viruses, bacteria and animals from worms to man. It was thought to be universal, but it is not: mitochondria have their own version, which is subject to variation in different organisms. Mammalian mitochondria read *AGA* and *AGG* as stop signals rather than as codons for the amino acid arginine and translate *AUA* as methionine rather than as isoleucine; *AUA* and *AUU* can serve as start signals instead of *AUG.* In the mitochondria of the molds *Neurospora* and *Aspergillus,* on the other hand, *AGA, AGG* and *AUA* are translated according to the "universal" code. In yeast, as in mammals, *AUA* codes for methionine, but the four codons beginning with *CU* are read as threonine rather than as leucine. In the mitochondria of mammals, yeast, *Neurospora* and *Aspergillus* the amino acid tryptophan is specified by *UGA* as well as by the usual *UGG,* but in maize the extra codon for tryptophan appears to be *CGG.*

There are 61 codons (4^3, minus the three stop codons), but organisms make do with fewer transfer RNA's because a single transfer RNA can read more than one codon. It can do so because there is some freedom (or "wobble") in the pairing of the third base of a codon with certain bases in the first position of an anticodon. Under the wobble rules deduced some years ago by F. H. C. Crick, 32 transfer RNA's (but no fewer) should be able to read 61 codons. Yet sequence analysis shows that human mitochondrial DNA codes for only 22 transfer RNA's, yeast mitochondrial DNA for 25. No additional transfer RNA's seem to be imported from the cytoplasm, and yet 61 codons are read. How?

Each of the 22 transfer RNA's in mammalian mitochondria appears to be able to read a "family" of either two or four synonymous codons [*see illustration on page 86*]. The transfer RNA's for two-codon families seem to have conventional *G*: *U* wobble anticodons, that is, anticodons beginning with a *G* that is able to pair with *U* as well as with *C* or beginning with a *U* that can pair with *G* as well as with *A.* The transfer RNA's for four-codon families, however, ignore the Crick rules: they apparently read all four synonymous codons. Their anticodons have a *U* in the 5′ position. Either that *U* somehow pairs with all four of the 3′ codon bases or the anticodons read only the first two bases and ignore the last one. The possibility of such a "two out of three" reading process was suggested several years ago by Ulf Lagerkvist of the Faculty of Medicine of the University of Göteborg.

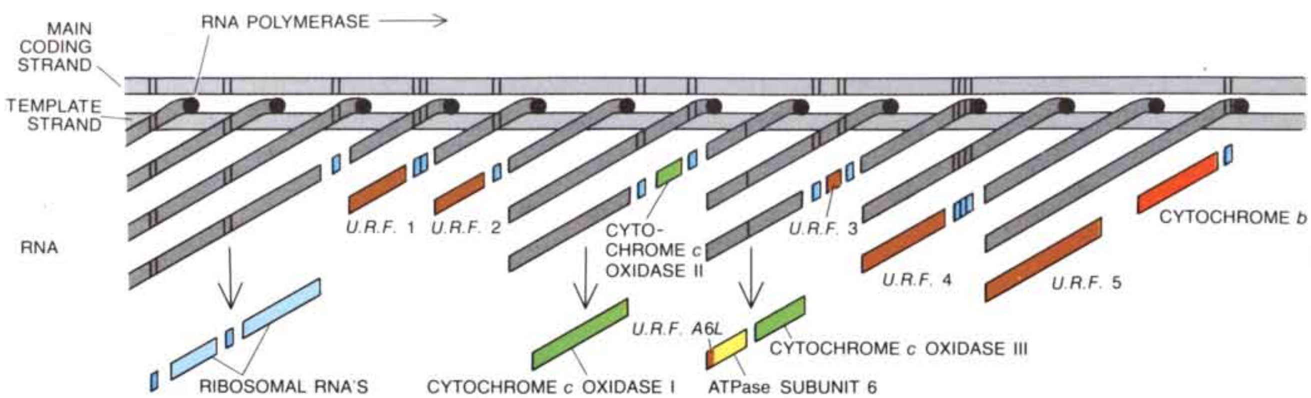
An intriguing question is how some transfer RNA's with *U* in the wobble position are held to reading two codons rather than four. What keeps them from recognizing codons ending in *U* or *C* as well as those ending in *G* or *A?* In the mitochondria of yeast and *Neurospora* and in many nonmitochondrial transfer RNA's such recognition is prevented by chemical modification of the wobble *U,* but so far sequence analysis of mitochondrial transfer RNA's in animals has not revealed any modification at this position. Most of these transfer RNA's do have an unconventional structure, however, and perhaps the structure influences the anticodon's specificity.

The Yeast Introns

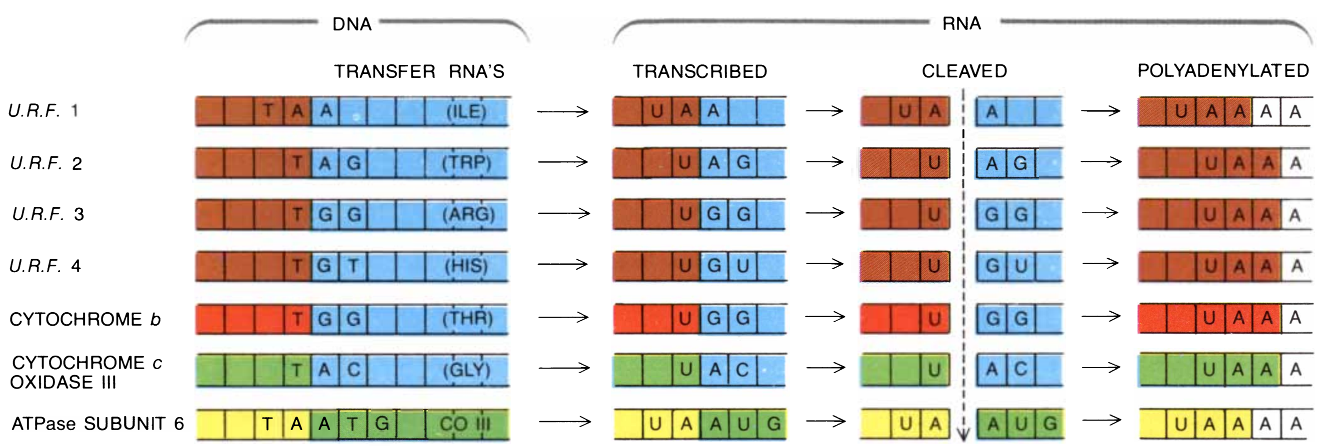
Split genes were first discovered in 1977, and by now it is clear that many of the chromosomal genes in mammals, birds and amphibians (and some of the genes in lower eukaryotic organisms) are split, with noncoding intervening sequences interrupting the "exons" coding for a protein. When such a gene is expressed, it is first transcribed, introns and all, into a precursor messenger RNA. This RNA is then "spliced," or processed: the introns are excised and the exons are joined to make a mature messenger RNA that is translated into protein. The introns of chromosomal genes, as far as anyone can tell, do not code for anything.

In yeast mitochondria things are different. Several years ago Piotr P. Slonimski and his collaborators at the Center for Molecular Genetics in Gif-sur-Yvette in France found that mutations in one intron of the yeast mitochondrial gene for cytochrome *b* (one of the respiratory enzymes) affect the splicing of the RNA transcribed from the gene. A series of elegant genetic experiments, coupled with DNA sequence analysis, revealed the reason: the splicing of the

SAMPLE OF THE NUCLEOTIDE SEQUENCE of human mitochondrial DNA is taken from the data gathered by Barrell and his colleagues. Only some 2,200 nucleotides of the 16,569-nucleotide molecule are shown here, but it is enough to suggest how tightly packed the genes are. In one region (beginning with nucleotide No. 8,527) two genes actually overlap, that is, the same nucleotides do double duty, encoding the end of U.R.F. *A6L* and the beginning of the gene for subunit 6 of ATPase. (The reading frame is different, as is indicated by the two brackets.) The sequence is the one for the strand that serves as the coding strand for most of the genes shown and that therefore has the same sequence as the corresponding messenger and transfer RNA's. The transfer RNA for the amino acid serine is actually specified by the complementary strand, which is transcribed in the opposite direction, as is indicated by its arrow.



HUMAN MITOCHONDRIAL DNA is transcribed into RNA and the RNA is simultaneously processed to form the mature ribosomal RNA's, transfer RNA's and messenger RNA's, according to Giuseppe Attardi of the California Institute of Technology. The diagram shows how the main coding strand is transcribed, beginning at a single major promoter site (*left*), into a single RNA transcript. The transcript is cleaved at the ends of the transfer-RNA sequences (*rectangles*) that flank almost every gene. The initial transcript and its successively processed segments are shown in gray. Mature RNA's are colored to match the scheme in the preceding three illustrations.



STOP CODONS signaling the termination of translation into protein are missing in seven human mitochondrial transcripts. In some the sequence runs into a downstream coding sequence; in others the three-nucleotide stop codon is cleaved, and thus truncated, when the RNA is processed. The necessary signal is supplied in an unconventional way by polyadenylation, a normal event in RNA processing in which a string of adenine nucleotides (*A*) is added to each transcript to make a messenger RNA. In this case the *A*'s fill out the stop codon *UAA*.

second intron of the gene is mediated by a protein that is specified in part by a segment of the intron itself!

Various forms of the cytochrome *b* gene have now been studied in different strains of yeast. The most complex form has six exons (designated *B*1 to *B*6) and five introns. Three of the introns include long unassigned reading frames. The entire gene is transcribed into a long primary transcript. Then intron No. 1 is excised. The effect of this step is to link exon *B*1 with exon *B*2 and at the same time to open up the reading frame in intron No. 2: to put its triplets in phase with those of the two preceding exons.

The spliced transcript of *B*1, *B*2 and the reading-frame portion of intron No. 2 now acts as a messenger RNA. It is translated into a protein whose amino-terminal head comprises the first 143 amino acids of cytochrome *b* (encoded by *B*1 and *B*2) and whose carboxy-terminal tail is encoded by the reading frame of the intron. This strange protein helps to catalyze the excision of intron No. 2. In so doing it destroys its own messenger RNA and thereby limits its own level in the mitochondrion. Slonimski calls this phenomenon "splicing homeostasis." He calls the protein encoded by the exon-intron combination a maturase to indicate that it may be not an enzyme itself able to catalyze splicing but rather a factor that somehow modifies the specificity of a preexisting splicing enzyme, probably one encoded by a nuclear gene.

Small Circular RNA's

The presence of an RNA maturase accounts neatly for the behavior of a number of mutant mitochondria having lesions within intron No. 2. They cannot process transcripts of the cytochrome *b* gene; instead they accumulate long RNA's from which only the first intron has been excised. They go on to generate large quantities of proteins that cross-react with antibodies to cytochrome *b* but are not cytochrome *b*. The synthesis of these proteins is blocked if mutations that terminate protein synthesis are introduced into exon *B*1. The proteins, then, must be defective maturases that arise from the translation of exons *B*1 and *B*2 and the long reading frame in intron No. 2; their overproduction is a direct result of their inability to destroy the messenger RNA that directed their synthesis.

Splicing homeostasis is probably not limited to the second intron of the cytochrome *b* gene. Introns No. 3 and No. 4 of that gene and the first four introns of the gene for subunit I of cytochrome *c* oxidase (another inner-membrane protein) also incorporate long open reading frames, and there is evidence that intron-specific maturases are involved in their processing.

Some mitochondrial splicing events generate unusual small circular RNA's, which were discovered by Annika C.

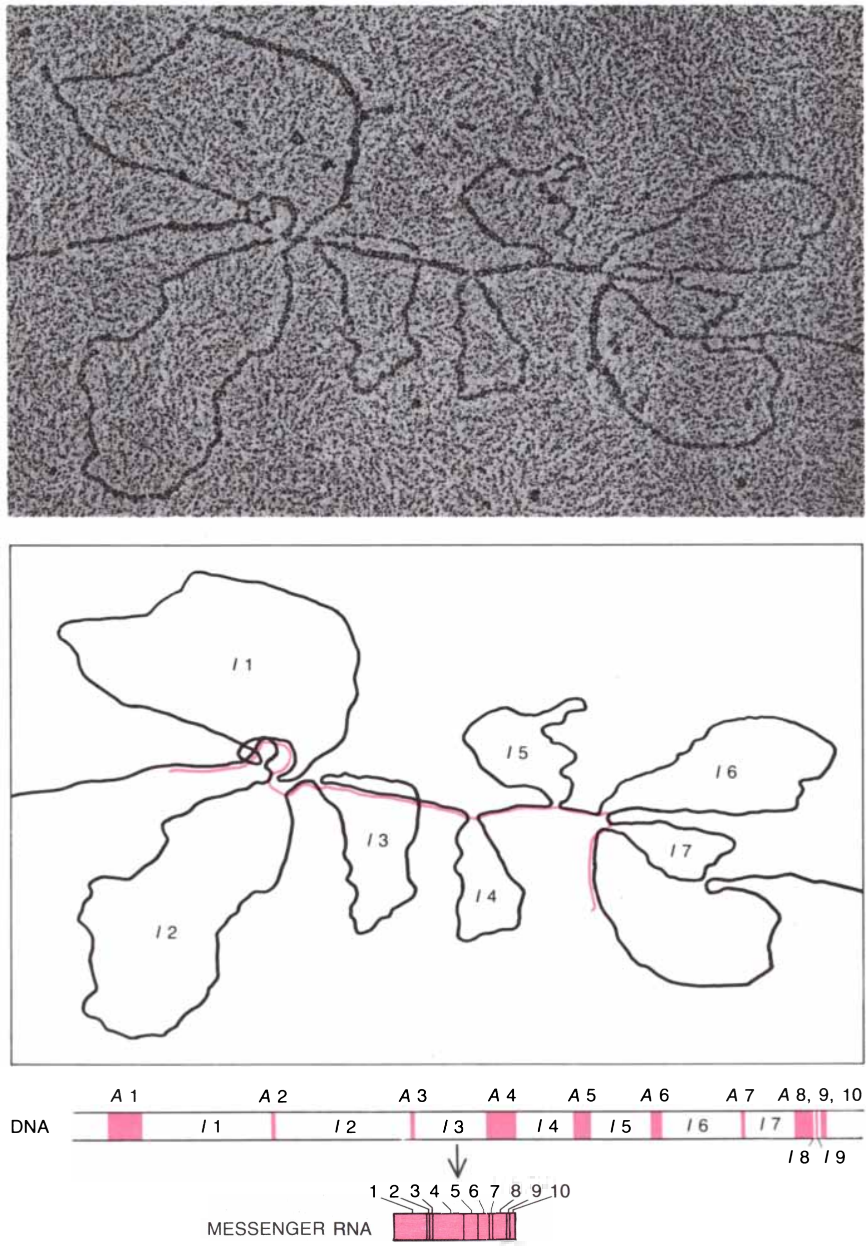
Arnberg of the State University at Groningen when she was collaborating with Gertjan van Ommen and me in electron-microscope studies to map mitochondrial transcripts. The circles turned out to be introns that had been spliced out of precursor RNA's and were then formed into closed loops, apparently by a covalent chemical bond. The circles have now been associated with three introns of the gene for subunit I of cytochrome *c* oxidase and with one intron (the No. 2 intron discussed above) of cytochrome *b*. The cytochrome *b* circle and one other from an intron with a long unassigned reading frame are apparently excised by mitochondrially encoded maturases, whereas the excision of the other two may depend on splicing enzymes encoded in the nucleus.

Crick once speculated that RNA circles might arise during splicing if a splicing enzyme can ligate the two ends of a single intron as well as join two exons. So far only one other instance has been reported where splicing gives rise to circular RNA's. In the nucleus of the protozoan *Tetrahymena,* Thomas R. Cech of the University of Colorado at Boulder and his colleagues reported a few years ago, an intron excised from the transcript of the gene for a ribosomal RNA is formed into a circle. When Cech's group tried to identify the enzyme that seemed to be mediating the excision and the circularization, they came to the surprising conclusion that there is no enzyme: the excision and circularization seem to depend on intrinsic properties of the RNA itself. The *Tetrahymena* RNA is the only example yet known of what Cech calls a ribozyme: a self-splicing RNA. Further study of mitochondrial circles that do not require maturases for excision should reveal whether ribozymes are implicated in their case as well.

Why Introns?

One remarkable property of the maturase associated with intron No. 2 is that it is dispensable. Yeast strains with shorter versions of the gene for cytochrome *b* are common; these genes often lack the first three introns, so that exons *B*1 through *B*4 are fused. The information originally present in the missing introns is not found elsewhere in the yeast mitochondrial genome, which suggests that the maturase is the mitochondrial equivalent of the worst kind of bureaucrat—one who exists only to regulate his own existence. Actually the three missing introns are members of a class of "optional" yeast mitochondrial introns defined by Johan Sanders when he was working in our laboratory at the University of Amsterdam a few years ago. There are other examples in the genes for the larger of the two ribosomal RNA's and for subunit I of cytochrome *c* oxidase. To classify an intron as optional one needs merely to find a strain that lacks it and yet is functional. Only a limited number of yeast strains have been examined so far; all yeast mitochondrial introns may eventually turn out to be optional. The optional nature of even a few introns puts some basic questions in a new light. Why are some genes split whereas others are not? Why should the same gene exist in both the split and the continuous form? Where do optional introns—and introns in general—come from?

Walter Gilbert of Harvard University



SPLIT GENE in yeast mitochondrial DNA is mapped by electron microscopy. The micrograph, made by Arnberg and Van Bruggen, and the interpretive map below it show a hybrid molecule: a single strand of DNA incorporating the gene for subunit I of cytochrome *c* oxidase (*black line in map*) has been allowed to hybridize with the messenger RNA (*color*) transcribed from it. In the DNA the sequences coding for the protein are interrupted by introns. Complementary sequences in the DNA and the RNA have paired to form double-strand regions; the introns, having no counterpart in the RNA, form single-strand loops. When such a gene is expressed, the primary RNA transcript is "spliced" to remove the introns, leaving only the "exons," or regions coding for the protein. Information from hybridization experiments has been combined with sequence data, gathered by Alexander Tzagoloff of Columbia University and by Lambert A. M. Hensgens in the author's laboratory, to reveal the organization of the exons (*A*1 through *A*10) and introns (*I*1 through *I*9) and of the mature messenger RNA after processing (*bottom*). (Introns *I*8 and *I*9 are too small to be distinguished in a micrograph.)

has argued that in nuclear and viral genomes introns may confer a selective advantage by promoting more rapid evolution. By greatly increasing the length of a gene they increase the likelihood of various recombination events in which analogous segments of two alleles, or alternative versions of the same gene, are interchanged. Such recombination might bring two advantageous mutations, each in a different allele, together in one allele to create a doubly advantageous gene. It might also create completely new genes by what Gilbert calls exon shuffling: by bringing together, as exons of a single gene, several coding sequences that had previously specified different proteins. Moreover, a single-base mutation at a boundary between an intron and an exon can change the splicing pattern and thereby generate messenger RNA's encoding very different proteins. The changes in splicing induced by such mutations need not be absolute, so that the affected DNA becomes a "multiple choice" gene: it can code for both the original protein and the novel one depending on alternative splicing events, enabling the organism to test a solution without erasing the old one from its memory.

How did introns arise? The introduction of irrelevant sequences in an essential gene is ordinarily lethal. Even if an RNA-processing mechanism excises such sequences from the primary transcript of a gene, why should selective forces preserve a newly introduced intron? The cell derives no immediate advantage from the presence of an intervening sequence and certainly cannot foresee that it may be advantageous in the future. Such considerations led W. Ford Doolittle III of Dalhousie University in Nova Scotia to propose that introns were not introduced into eukaryotic cells but rather are part of their evolutionary heritage.

In Doolittle's view present-day prokaryotes (the bacteria) as well as the eukaryotic nucleus are likely to be derived from a primitive ancestral organism that was able to evolve rapidly because it had sloppy, unfaithful machinery for DNA replication and transcription and for translation. Such an organism's genetic information would have to have been encoded redundantly if any of it was to be expressed correctly. A genes-in-pieces organization might have been essential to ensure that transcripts of exons were at least occasionally assembled to direct the synthesis of functional proteins. As the genetic machinery became more accurate, Doolittle suggests, such insurance became less necessary; the replication and transcription of redundant and noninformational DNA became increasingly burdensome and there was pressure to eliminate it. The advantage of elimination was increased efficiency as exemplified by the streamlined genes of modern bacteria and the viruses that infect them. The price paid for this increased efficiency was the loss of a certain potential for further evolutionary change.

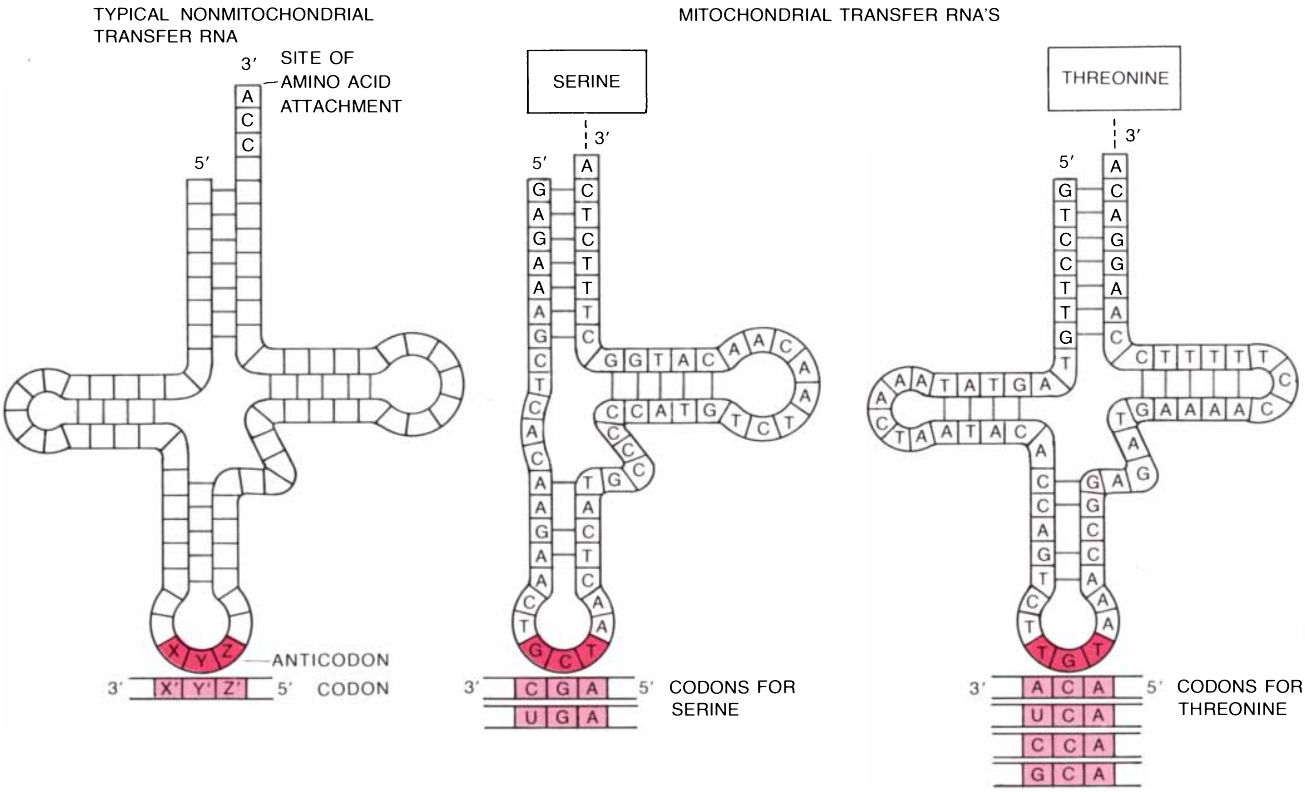
What about introns in mitochondrial genes? Many of the proposals advanced to explain introns in nuclear genes are unsatisfactory for mitochondrial ones, not only because they fail to explain the unassigned reading frames but also because their main predictions are not supported by the facts. Passive retention of introns with unassigned reading frames is not a good explanation because the cost to the cell of maintaining and expressing the many genes required to splice introns is appreciable, and such a complex system must be very vulnerable to disruption by mutation. Retention to facilitate recombination is an unlikely explanation because there is no indication that new mitochondrial genes have arisen recently by exon shuffling or that mitochondrially encoded proteins in fungi are evolving any faster than their counterparts in mammalian mitochondria, whose genes lack introns. Moreover, although exon shuffling is in principle an excellent way to spread mutations conferring advantageous characteristics (such as resistance to naturally occurring inhibitors of mitochondrial function), there is no reason to believe such mutations are fixed more efficiently in split genes than they are in continuous ones. Finally, there is no indication (although solid data are sparse) that any of the split genes in yeast mitochondria are multiple-choice genes.

Introns and Repression

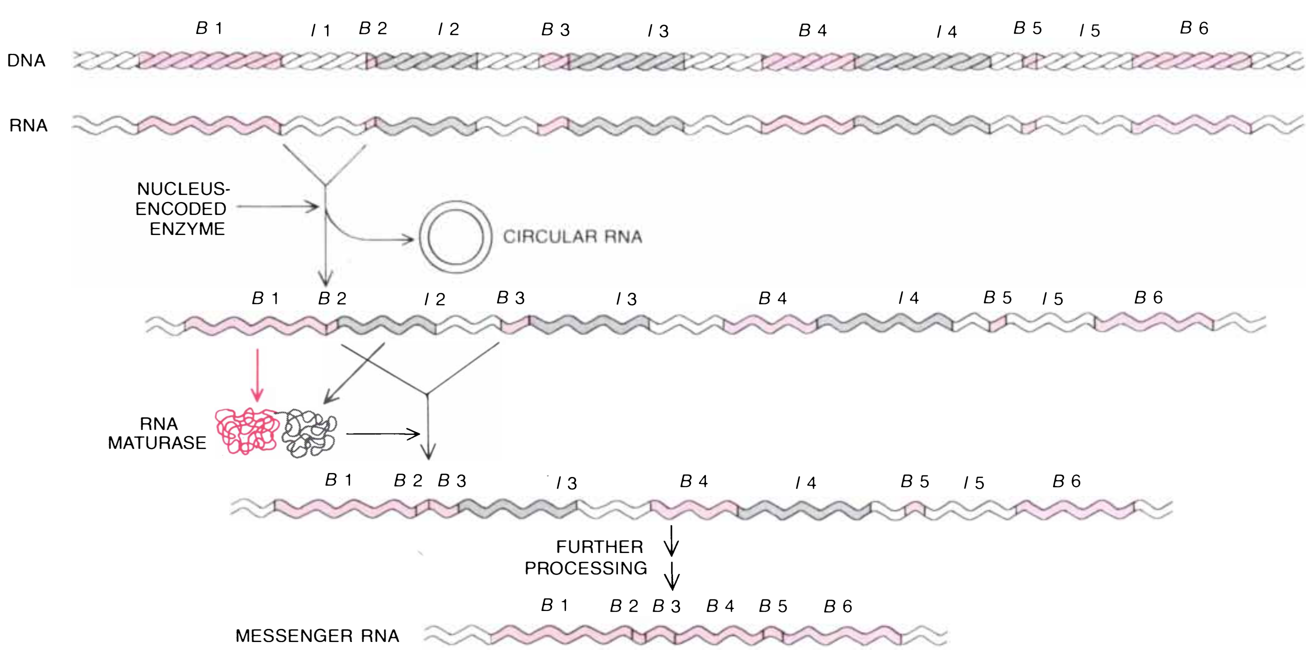
Perhaps, then, introns with unassigned reading frames have been retained in fungal mitochondrial DNA's because they confer some kind of selective advantage. What could the advantage be? One plausible explanation, put forward by Piet Borst of the University of Amsterdam, is that such introns may have a role in the fine regulation of mitochondrial-gene expression. Yeast cells can selectively repress the synthesis of certain mitochondrial proteins when the cells are growing by fermentation in the absence of oxygen, and some of the repression may be achieved at the level of RNA processing. Borst's proposal has the advantage of explaining why the repression is selective. Such genes as those for cytochrome *b* and subunit I of cytochrome *c* oxidase, which have several introns whose excision is likely to be controlled by a maturase, should be particularly sensitive to repression—and this has been found to be the case. The repression argument does not explain the presence of all yeast mitochondrial introns, however. One should not overlook the possibility that at least some of the introns without unassigned reading frames may not have any function at all. Given an effective mechanism for removing them from RNA, there may be little selective pressure to eliminate them from the DNA; indeed, mitochon-

FIRST LETTER	SECOND LETTER				THIRD LETTER
	U	C	A	G	
U	PHE	SER	TYR	CYS	U
	PHE	SER	TYR	CYS	C
	LEU	SER	*STOP*	*(STOP)* TRP	A
	LEU	SER	*STOP*	TRP	G
C	LEU	PRO	HIS	ARG	U
	LEU	PRO	HIS	ARG	C
	LEU	PRO	GLN	ARG	A
	LEU	PRO	GLN	ARG	G
A	ILE (MET)	THR	ASN	SER	U
	ILE	THR	ASN	SER	C
	(ILE) MET	THR	LYS	(ARG) *STOP*	A
	MET	THR	LYS	(ARG) *STOP*	G
G	VAL	ALA	ASP	GLY	U
	VAL	ALA	ASP	GLY	C
	VAL	ALA	GLU	GLY	A
	VAL	ALA	GLU	GLY	G

GENETIC CODE, in which a triplet of nucleotides (designated by their initials *U, C, A* and *G*) serves as a codon specifying a particular amino acid or a stop signal, was thought to be universal but is actually different in mitochondrial genetic systems. The table shows how codon assignments deviate in human mitochondria from those of the "universal" code; each mitochondrial deviation is shown in color, with the normal translation in brackets. Each of the 22 transfer RNA's in human mitochondria "reads" either two synonymous codons (*gray boxes*) or four (*colored boxes*). The codon response of the methionine transfer RNA's is not yet clear.



TRANSFER RNA's generally have a stereotyped cloverleaf structure (*left*), the result of pairing (*gray lines*) between complementary nucleotide bases. (In RNA *A* pairs with *U*, and *G* pairs with *C*. In the mitochondrial transfer RNA's illustrated here the sequence is that of DNA rather than RNA, and so *A* is seen to pair with *T*.) Many mitochondrial transfer RNA's have an aberrant shape; the one for serine (*middle*) lacks a complete arm. A three-nucleotide anticodon on each transfer RNA reads a codon of messenger RNA. In some cases a transfer RNA can read more than one codon because there is some freedom, or "wobble," in the pairing between the first base of its anticodon and the third base of a codon; the *X* of the anticodon at the left might pair with a base other than *X′*. Many transfer RNA's in human mitochondria can apparently read as many as four codons (*right*), so that only 22 transfer RNA's suffice to read 61 codons.



RNA MATURASE, an enzyme specified in part by the second intron in the yeast mitochondrial gene for cytochrome *b*, helps to catalyze the excision of the intron encoding it. The gene has six exons (*B*1 through *B*6) and five introns. (Exons and introns are not drawn at the same scale.) It is transcribed into a precursor RNA, from which the first intron is promptly excised by a nucleus-encoded splicing enzyme. (Copies of the intron accumulate, surprisingly, in the form of single-strand circles.) The excision joins exons *B*1 and *B*2 to a long unassigned reading frame (*gray*) within intron *I*2. The RNA so assembled serves as a messenger RNA specifying the maturase, which is necessary for excision of *I*2; the maturase thus controls its own synthesis. There is evidence that other maturases help to excise *I*3 and *I*4.

dria may lack the machinery to do so efficiently. The introns could be a form of "selfish DNA," perpetuating itself for its own sake.

The fact that some introns in yeast, *Aspergillus* and *Neurospora* mitochondria are at closely corresponding sites in the genes they split suggests that the common ancestors of the genes may have had a similar organization. At least these introns may therefore directly reflect the genes-in-pieces organization of the primitive endosymbiotic organism from which mitochondria are thought to have evolved. Not all introns may be that old, however. Detailed comparisons of intron sequences in yeast, done by François Michel and Bernard Dujon of the Center for Molecular Genetics and by Linda Bonen of the University of Amsterdam, reveal many similarities both in the secondary structure of the RNA and in protein-coding sequences. These similarities strongly suggest that most present-day introns and their unassigned reading frames are members of a multigene family that originated in a single ancestral gene and has expanded by duplication and transposition in the course of evolution.

Such duplication with mobility is characteristic of the behavior of transposons: pieces of DNA that, in bacteria and in higher organisms, replicate and insert their copies at random sites in a genome. The closely similar yeast introns may possibly have arisen from transposons. If that is the case, it could explain the origin of the maturases. Most bacterial transposons encode a transposase: an enzyme that catalyzes the insertion of the transposon. Transposition involves cleavage precisely at the two ends of the transposon, just as RNA splicing requires precise cuts at the ends of an intron. The maturases could, therefore, have evolved from transposases.

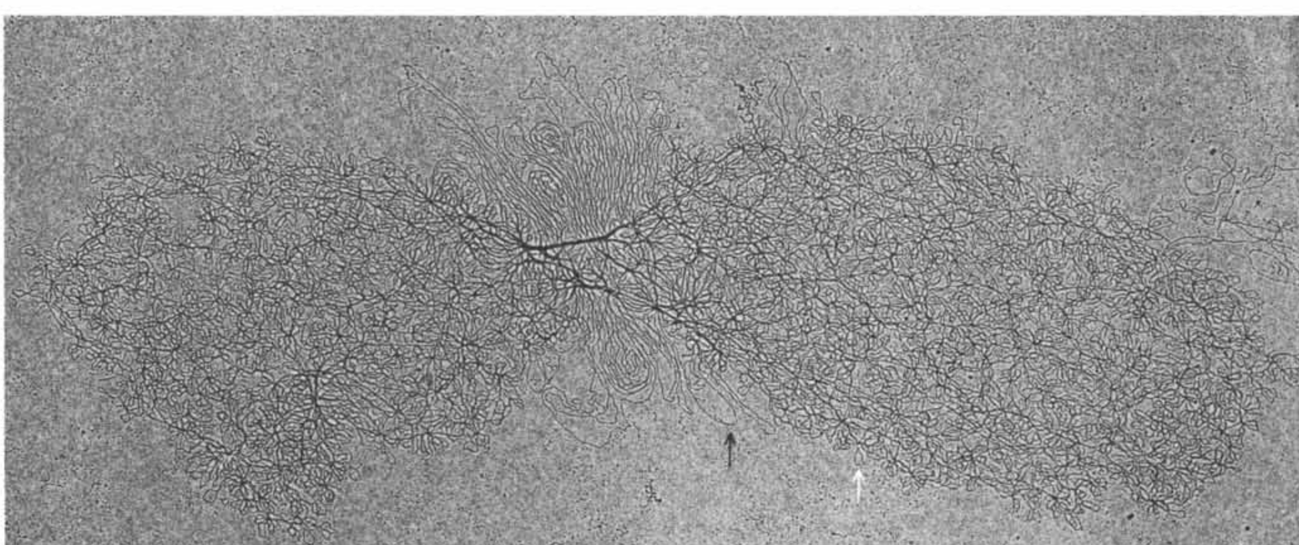
Why Mitochondrial DNA?

Why should a separate genetic system be necessary for just one cellular organelle among many? The mitochondrial system cannot be explained as a form of gene amplification, necessary to provide the mitochondrion with large quantities of certain proteins. High levels of protein synthesis can be achieved by other means, and if amplification were necessary, it could as well take place in the nucleus. Nor can mitochondrial DNA be explained as the necessary source of proteins that are too insoluble to be transported to their site in the inner membrane. Subunit 9 of the ATPase complex, one of the most hydrophobic of the proteins specified in yeast by mitochondrial DNA, is specified by a nuclear gene in *Neurospora* and in mammals, synthesized on ribosomes in the cytoplasm and imported into the mitochondria.

If a separate mitochondrial genetic system is not necessary, why is it there? Borst thinks mitochondrial DNA is simply an evolutionary dead end, a relic of the putative symbiotic bacterium. According to Borst's hypothesis, in the early stages of evolution there were few barriers to the transfer of information between the genetic system of the endosymbiont and that of the host, and gradually most of the endosymbiont's genes became part of the host cell's genome. As the eukaryotic cell became more sophisticated (as its DNA was organized in chromosomes and segregated in a nucleus) and perhaps as the nuclear and the mitochondrial genetic codes diverged, the likelihood of transfer decreased; finally it became impossible. A few genes remained isolated in the mitochondrion. Being essential, they persist.

The hypothesis is attractively simple. It accounts nicely for the fact that a number of mitochondrial enzymes encoded in the nucleus have similarities to corresponding enzymes in bacteria. It predicts that some genes will be found in the mitochondrial DNA of one organism but in the nuclear DNA of another, as in the case of the gene for subunit 9 of ATPase. It allows for intermediate situations in which both mitochondrial and nuclear versions of the same gene are present in the same cell. That may be true in *Neurospora,* in which Paul van den Boogaart and Étienne Agsteribbe of the State University at Groningen have found copies of the gene for subunit 9 of ATPase in both mitochondrial and nuclear DNA.

Borst's hypothesis predicts further that some mitochondrial DNA's may have more genes than are in the basic set present in both yeast and mammalian mitochondria. The mitochondria of higher plants have extremely large genomes, ranging from perhaps 250,000 nucleotides in maize to as many as 2.4 million in the squash and melon family. Much of the difference may be due to the presence of noncoding DNA, but laboratory experiments show that plant mitochondria synthesize more proteins than their fungal or mammalian counterparts do, and there is good evidence that in plants the phenomenon of male



UNUSUAL MITOCHONDRIAL DNA of the protozoan *Trypanosoma brucei* is enlarged some 16,000 diameters in an electron micrograph made by J. H. Hoeimakers of the University of Amsterdam. The DNA is in the act of segregation: having been replicated, it is dividing to form two genomes when the organelle divides. The DNA is a network of interlocked maxicircles (*black arrow*) and minicircles (*white arrow*). The maxicircles seem to be equivalent to mitochondrial DNA in other organisms; function of minicircles is not known.

sterility (inability to produce fertile pollen) can be the result of changes in mitochondrial DNA.

Evolution toward Diversity

Whereas the major products of mitochondrial genes appear to have been largely conserved in the course of evolution, the organization and the mode of expression of those genes have not been conserved. The diversity of organization is nowhere better exemplified than in the mitochondrial DNA of *Trypanosoma,* a genus of unicellular protozoans that includes parasites of man and domestic animals. The DNA is an enormous network of linked circles—perhaps 10,000 of them. There are maxicircles ranging in length from some 20,000 to 36,000 nucleotides. They have regions of homology with pieces of yeast DNA and are transcribed; they are likely to be the counterpart of the mitochondrial DNA in other organisms. Most of the DNA, however, consists of minicircles less than 3,000 nucleotides long. These are not transcribed, and they have undergone evolution at a high rate, which is characteristic of noncoding DNA. The function of the vast network of minicircles is not known, nor is it at all clear why trypanosomal mitochondrial DNA should have such an unusual structure.

This final visit to one of the outposts of biology will, I hope, convince the reader that mitochondrial genetic systems are diverse. Their variety is the result of a high rate of nucleotide-sequence evolution—perhaps upward of 10 times the rate for genes in the nucleus. The apparent lack of restraint on sequence evolution may be related to the simplicity of the system. Its function is primarily to synthesize a limited number of proteins in rather constant amounts, and it may be able to accommodate mutations that would be deleterious in a system geared to the differential regulation of many genes. It is possible too that in each organism different selective pressures have pushed the mitochondrial DNA's into different and sometimes extreme solutions to the problem of gene expression.

Finding out what these pressures may be is one task for the future. Another is to learn through study of this limited system just what the minimal requirements of a functional genetic system are. Better understanding of the mitochondrial genome should reveal some of the evolutionary pathways leading to mechanisms of gene expression in present-day bacteria and eukaryotes. To judge from what the study of mitochondria has already revealed, the lessons to be learned about the organization and the expression of mitochondrial DNA should have applications extending far beyond the boundary membrane of this modest organelle.

The Future of the Universe

A forecast for the expanding universe through the year 10^{100}: All protons will decay, galaxies will form black holes and black holes will "evaporate." If the universe collapses, it may cycle

by Duane A. Dicus, John R. Letaw, Doris C. Teplitz and Vigdor L. Teplitz

In the past several years developments in the physics of high-energy interactions of elementary particles have stimulated a remarkable growth in cosmology. The attempt to describe all the fundamental forces of nature as different aspects of one underlying force has had partial success in the "grand unified theories" of elementary particles. Through such theories it is possible to give tentative descriptions of dominant physical processes from extremely low temperatures to temperatures on the order of 10^{32} degrees Kelvin. The properties of matter at densities of cosmological interest, which vary from less than 10^{-300} gram to more than 10^{100} grams per cubic centimeter, can thereby be addressed. The extreme conditions at the ends of these scales can prevail only in the very early or the very late stages in the evolution of the universe.

A few years ago cosmologists began to show how theoretical physics can give a consistent account of the history of the universe to within 10^{-35} second of its beginning. More recently several physicists and cosmologists, including the four of us, have extrapolated cosmic events into a future that is up to 10^{100} times the current age of the universe. Apart from the intrinsic fascination of such an enterprise, there are several reasons for undertaking it. From the standpoint of theoretical physics the extrapolation makes possible a kind of thought experiment: By determining how the diverse effects predicted by a theory interact at a given time, one can test the overall consistency and plausibility of the theory. No ordinary laboratory is ever likely to attain the temperatures and densities at which grand unified theories make predictions, and so the theories are being verified in what some physicists call the ultimate laboratory, namely the universe as a whole. From the standpoint of cosmology the importance of our extrapolations is that the grand unified theories also have consequences that can be tested in terrestrial laboratories, and so the predictions they make under extreme conditions can be confirmed. One can therefore delineate a sequence of events in the distant future with considerably more detail than has previously been possible.

The framework for our calculations about the remote future is the big-bang model. There is now general agreement that the evolution of the universe must have been determined in the first few moments of the big bang. According to the model, the universe began as the explosion of a highly compact, primordial entity some 10 to 20 billion years ago. The term big bang and the terminology associated with an explosion are appropriate in part, because matter and energy do indeed appear to have been flung outward in space from a common origin. The term can also be misleading, however, because it is not correct to suppose the expansion can be visualized from the outside. There is no external vantage because what is exploding is the entire universe. Space itself is expanding in the sense that the separation between any two galaxies is growing at a rate that depends on their separation. Viewed from our galaxy all other galaxies appear to be receding at a rate that increases by 17 kilometers per second for every million light-years' separation. Mathematically the big bang is described by Einstein's general theory of relativity.

The recession of distant galaxies is inferred from the observed shift of their radiation toward the red end of the electromagnetic spectrum. The expansion and cooling of the universe has altered even more dramatically the spectrum of another kind of radiation, namely the cosmic microwave background radiation, discovered in 1964 by Arno A. Penzias and Robert W. Wilson of Bell Laboratories. The background radiation seems to be coming with equal intensity from every point in the sky; the intensity varies with direction by less than one part in 10,000. It was the scattering and annihilation of electrically charged particles in the early stages of the big bang that gave rise to the background radiation. The radiation was therefore once very hot, but it has since been cooled by the expansion of the universe to a temperature of about three degrees K.

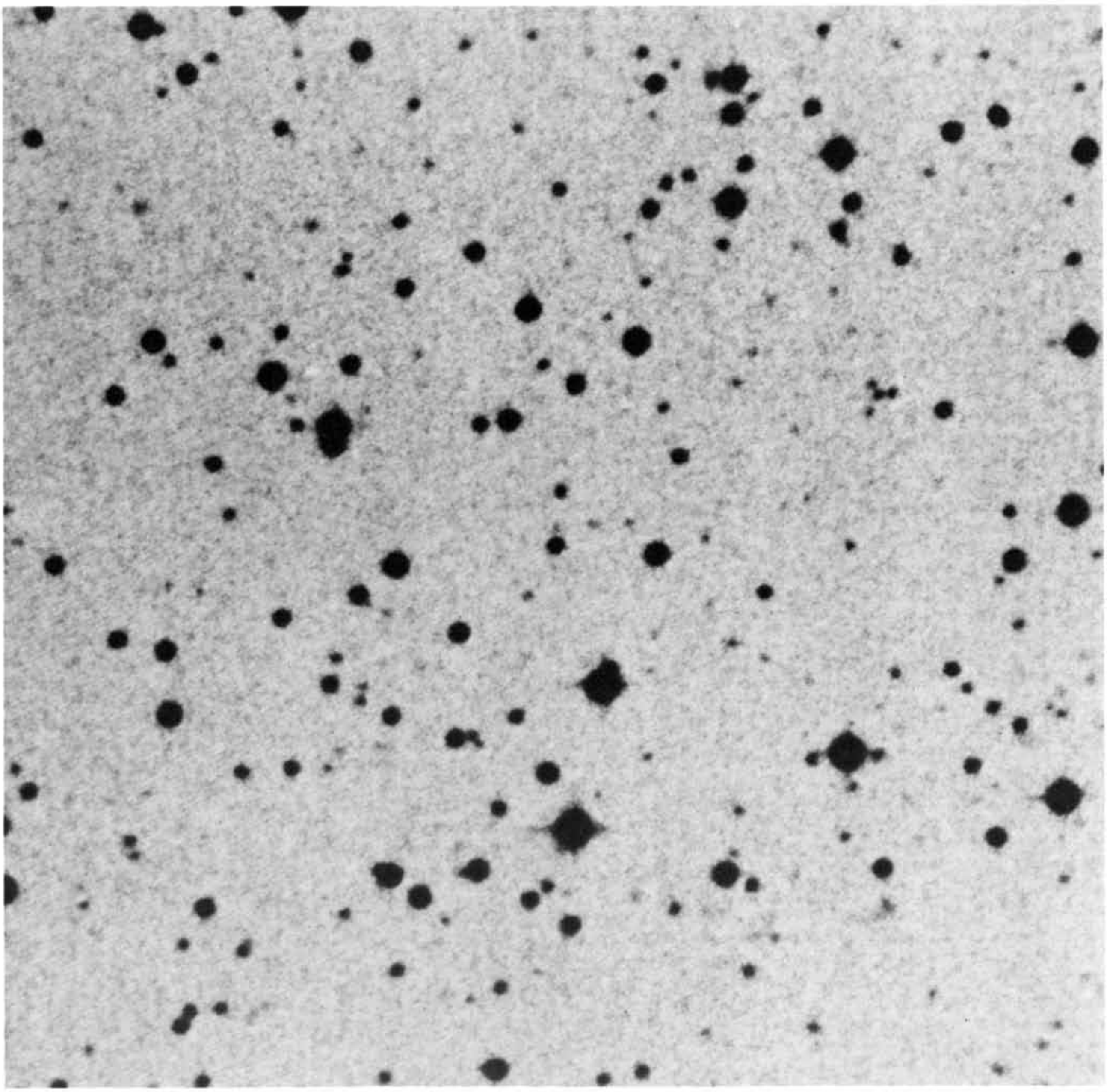
Nuclear reactions at the density and temperature of the universe about three minutes after the start of the big bang were responsible for the synthesis of helium and, to a much lesser extent, other light elements. The universe cooled too quickly, however, for carbon and heavier elements to form, and a generous supply of hydrogen was left over to serve as nuclear fuel for stars. Earlier, reactions only 10^{-38} second after the universe began may have generated the observed excess of matter over antimatter, in which case most grand unified theories predict the eventual decay of all nuclear matter. The current understanding of the conditions in the early universe has not, however, been sufficient to answer the most compelling question in cosmology: Will the universe continue to expand forever, or will gravitational forces halt the expansion and drag all space and time back into the fireball with which the universe began?

Although many years of observational effort and theoretical consideration have been given to the last question, the issue is still undecided. In principle the simplest way to resolve it is to measure the amount of matter in a large, representative volume of space and so estimate the density of matter in the universe at large. If the density is equal to or greater than a critical density, which is about three protons for every 1,000 liters of space, gravity will eventually overcome the expansion and the universe will collapse. A universe of this kind is said to be closed. If the density is less than the critical value, the universe will remain open; the expansion will be slowed by gravity but never stopped. Note that physical processes such as stellar burning or the decay of elementary particles cannot change a closed universe into an open one.

Estimates of the number of stars in a typical galaxy and the number of galaxies per unit volume give a value for the density of matter that is only 5 percent of the critical density. On this basis one might conclude that the universe is open and will continue its expansion forever. An estimate of mass based on the counting of stars, however, includes only luminous matter; if nonluminous matter accounts for a significant fraction of the mass of the universe, the estimate is not accurate.

There is some evidence of nonluminous matter in galaxies. The speed with which a star revolves about the center of its galaxy should increase with distance from the galactic center for stars lying in the interior of the galaxy and should decrease with distance if the star is at the periphery. For outlying stars, however, the expected decrease in speed of revolution is not observed; the observed distribution of speeds leads to the conclusion that at the periphery of a galaxy there is a halo of nonluminous matter whose total mass is one or two times that of the luminous galactic matter. Hence nonluminous matter within galaxies may account for from 5 to 10 percent of the mass needed to close the universe.

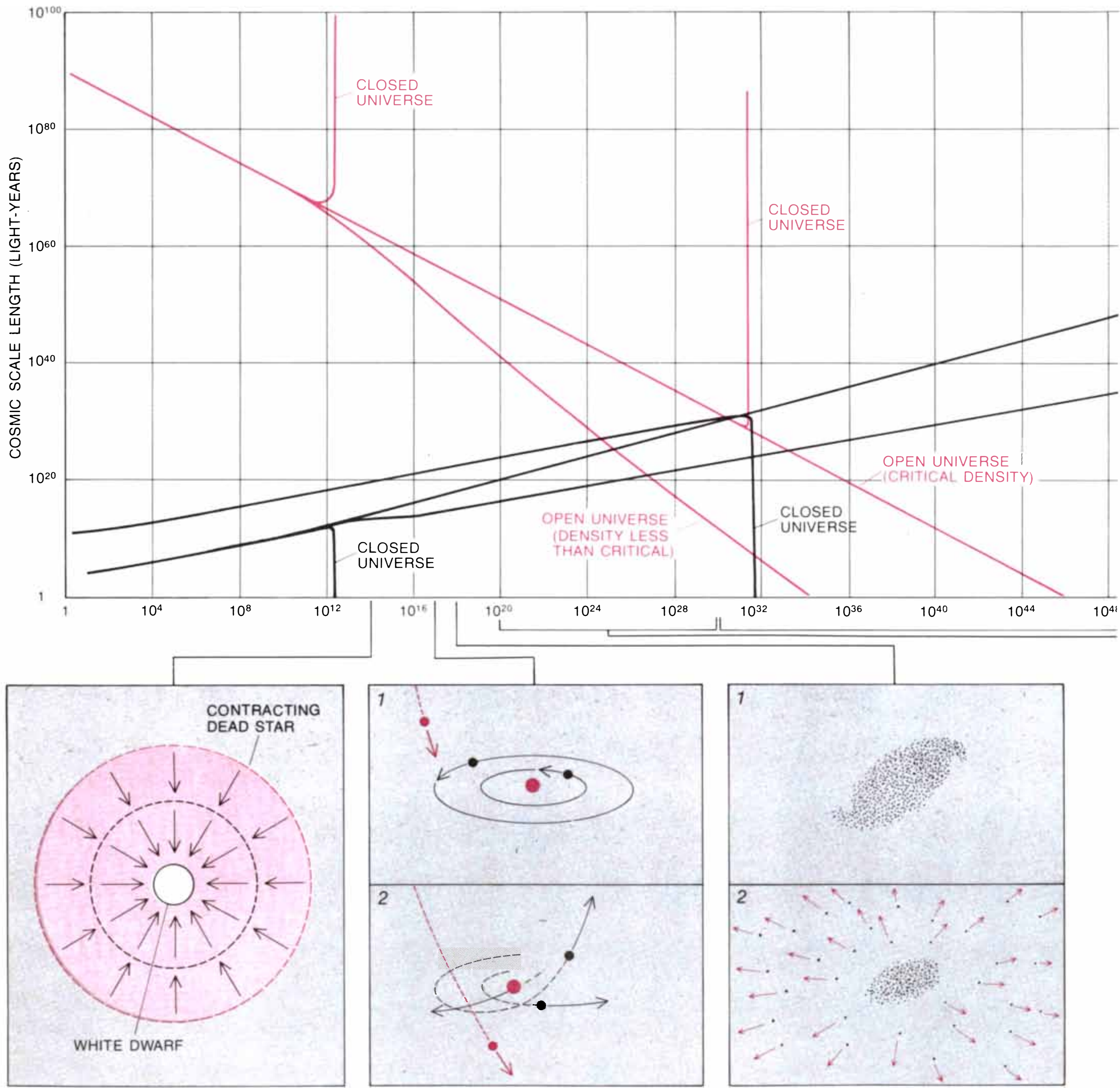
Similarly, the observed motions of galaxies in clusters of galaxies appear to require enough nonluminous matter to make up 50 percent of the critical density of the universe. The nonluminous matter would explain why the galax-



MOST DISTANT OBJECT KNOWN is a quasar designated PKS 2000–330; it has a red shift of 3.78, which implies it is receding from us at 92 percent of the speed of light. The enormous recessional velocity is probably an effect of the big bang; if it is, PKS 2000–330 is more than 16 billion light-years from our galaxy. The photograph is a negative print; the quasar is the dark, starlike object in the exact center. Such distant objects enable cosmologists to infer some of the conditions that prevailed in the distant past and so to extrapolate certain processes into the remote future. Light now reaching the earth from the quasar must have been emitted shortly after the big bang.

ies remain bound in clusters instead of moving off in random directions. The dark matter could be in the form of cold rocks, burned-out stars, clouds of gas or even black holes.

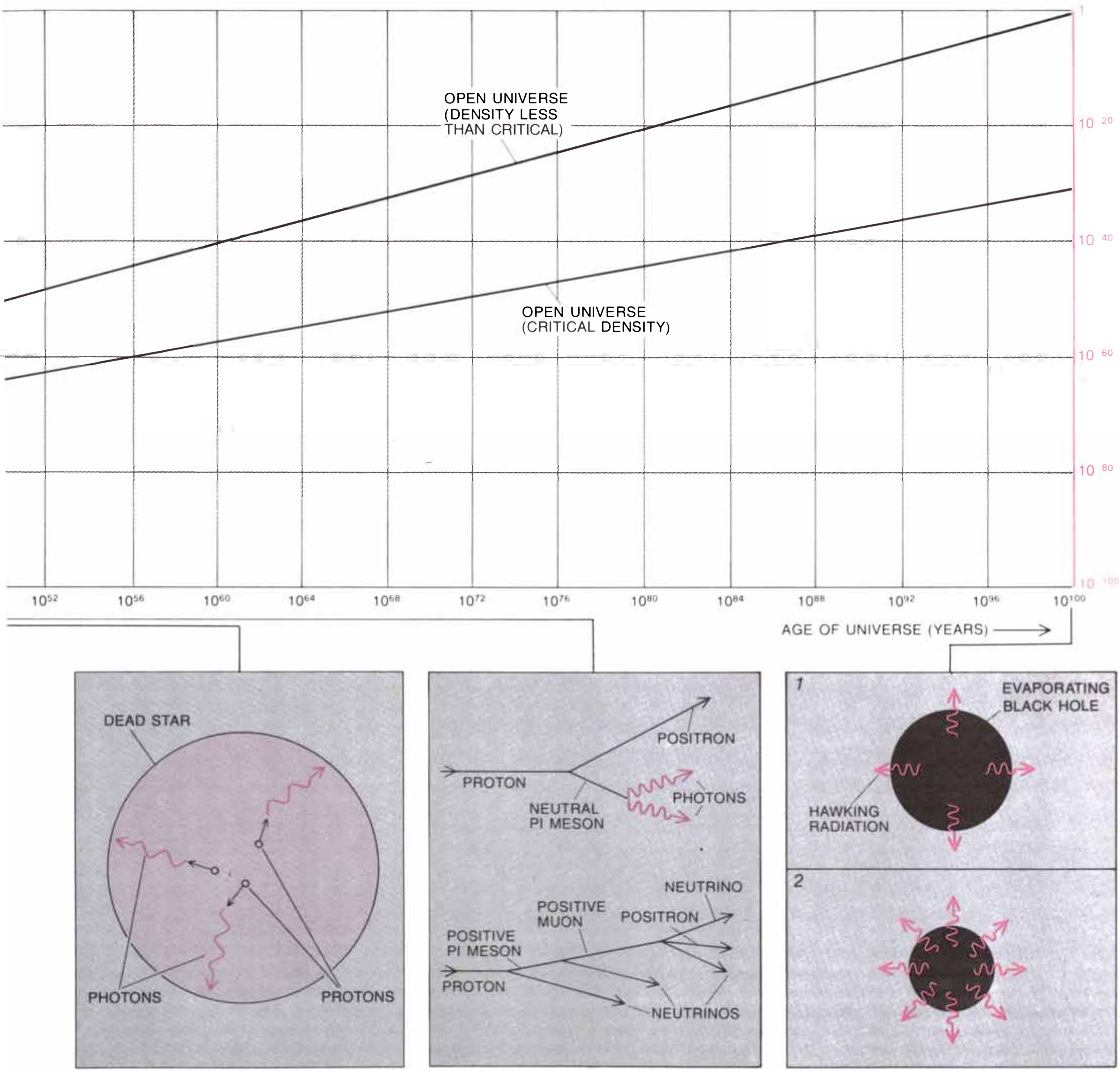
If the additional matter necessary to make up the remaining 50 percent of the critical density actually exists, it may consist of ionized hydrogen atoms or of neutrinos. Ionized hydrogen should be detectable from X-ray telescopes in orbit around the earth. As for neutrinos, there should be about as many of them as there are photons, or quanta of electromagnetic energy, associated with the cosmic background radiation; this amounts to a few hundred neutrinos per cubic centimeter. Their contribution to the density of the universe depends on the mass of the neutrino. It was once supposed the neutrino is massless, but current experiments only place upper bounds on neutrino masses and do not rule out the possibility that neutrino masses are sufficient to close the universe. Moreover, certain recent theoretical arguments suggest that the neutrino has some mass, and several groups of experimenters are trying to detect and measure it. Such experiments could definitely answer the question of closure in the affirmative within the next few years. A negative result, on the other hand,



SIX MAJOR EVENTS in the future of the universe are expected on the basis of physical theory, provided the expansion of the universe initiated by the big bang continues long enough for them to take place. Here the scale length (which is roughly equivalent to the radius of the universe) and the density are plotted against time for four models of the universe: one in which the universe is open, two in which it is closed and one on the mathematically defined boundary between the closed and the fully open models. If the universe is open, the density of matter and energy is too low for gravity to overcome the expansion, and so the universe will expand forever. If the universe is closed, gravity will eventually halt the expansion and drag the universe back into a point of infinite density. It is not known whether we inhabit an open or a closed universe. The future events are illustrated schematically: the stars run out of fuel and contract would leave the question to be resolved by more precise measurement of the distances and recessional velocities of remote galaxies.

Although the current uncertainty about the density of the universe straddles the value that distinguishes a closed universe from an open one, the uncertainty is confined to less than two orders of magnitude. The fact that the universe has not contracted before reaching its present age implies that the density is not greater than about 10 times the critical density. The estimate of density based on luminous matter alone indicates that the actual density must be at least a twentieth of the critical density. No one yet knows how to account for what cosmologists find to be a striking coincidence.

Since the question of closure has not yet been settled on experimental or observational grounds, both the open and the closed possibilities need to be considered in any speculations about the distant future. Suppose, to begin, that the critical density is not attained and the universe is open. What is likely to happen? The question must be answered both for large-scale structure (that is, for the evolution of the geometric properties of the universe) and for local com-

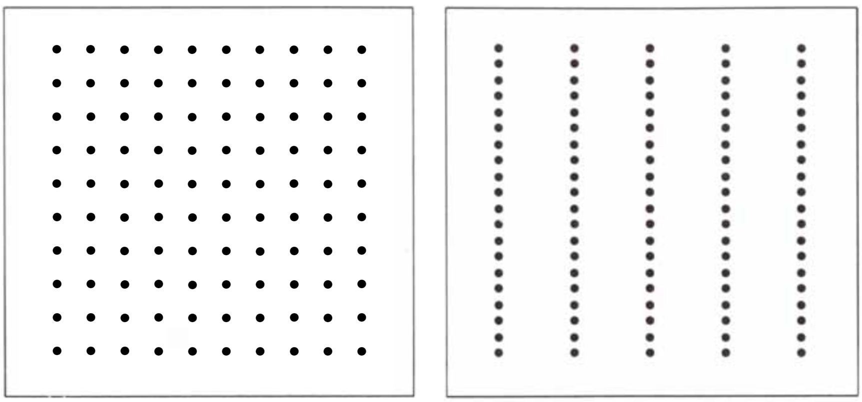


under their own weight in 10^{14} years; they lose their planets through close encounters with other stars in 10^{17} years. Galaxies lose a large fraction of their mass in about 10^{18} years because of still closer encounters between stars and because of scattering of both hydrogen atoms and dust particles; aggregates of matter that acquire a large velocity through the encounters can overcome the gravitational forces within the galaxy, and the remaining matter collapses into a supermassive black hole. (A black hole is a region of space from which nothing can escape except through mechanisms described by the quantum theory.) The decay protons generate heat inside the cold stars that escaped from the galaxies. The heating becomes significant in 10^{20} years, and by the year 10^{30} some 40 percent of the matter in the universe has undergone such decay. After 10^{100} years the supermassive black holes lose their mass through quantum "evaporation."

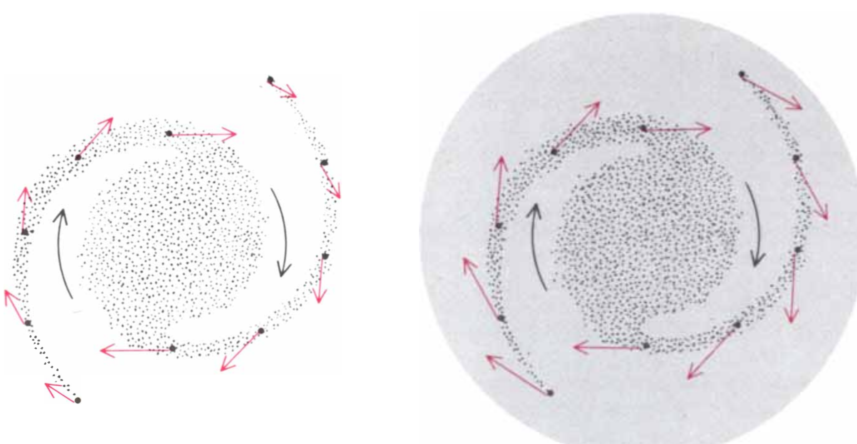
position (for aggregates of matter that range in size from protons to galaxies).

According to our best understanding, the local composition of the open universe will undergo six major transitions. First, within 10^{14} years after the big bang all stars will have run out of fuel. The principal nuclear fuel is hydrogen, which is converted into helium in the core of a star throughout most of its lifetime. After most of the hydrogen fuel has been consumed the star rapidly swells to many times its former size and becomes a red giant; in this stage of the star's existence helium is usually converted into carbon and heavier elements. These thermonuclear reactions are effectively one-way transitions: hydrogen is converted into helium, helium into carbon and carbon into heavier elements in a series that generally terminates with iron. The iron nucleus has a lower total energy per unit of mass than any other nucleus, so that once the "iron limit" has been reached the energy of the universe that is stored as nuclear fuel will have been entirely released.

The rate at which nuclear fuel is consumed by a star depends on the mass of the star: the more massive the star is, the faster it burns and so the shorter its lifetime is. The sun, for example, will use up much of its hydrogen within about 10 billion years, whereupon it will rapidly fluctuate in size, burn through some of the heavier elements at a prodigious rate but for a relatively short time and finally collapse into a small, slowly cooling white-dwarf star. The smallest stars will require many times the life span of the sun to pass through all these stages, but eventually they too will reach the iron limit. Note that although the exhaustion of nuclear fuel is the first major event in the future of the open universe, it will happen at a time exceedingly remote from our own. The last stars will cease to shine only after the universe is 10,000 times as old as it is now.



DISTRIBUTION OF MATTER AND ENERGY in the universe appears to be both homogeneous and isotropic to a high degree of precision. In a homogeneous universe equal numbers of galaxies fill comparable volumes of space. In an isotropic universe the density of matter and energy is the same in all directions as well as being homogeneous. The two properties are not equivalent: a homogeneous universe need not be isotropic. If matter is distributed as it is in the schematic diagram at the left, the universe is both homogeneous and isotropic. The distribution shown at the right is homogeneous but not isotropic. The distance between galaxies along the vertical direction is less than the distance along the horizontal direction. Anisotropy in an expanding, homogeneous universe is created by an expansion rate that varies with direction.



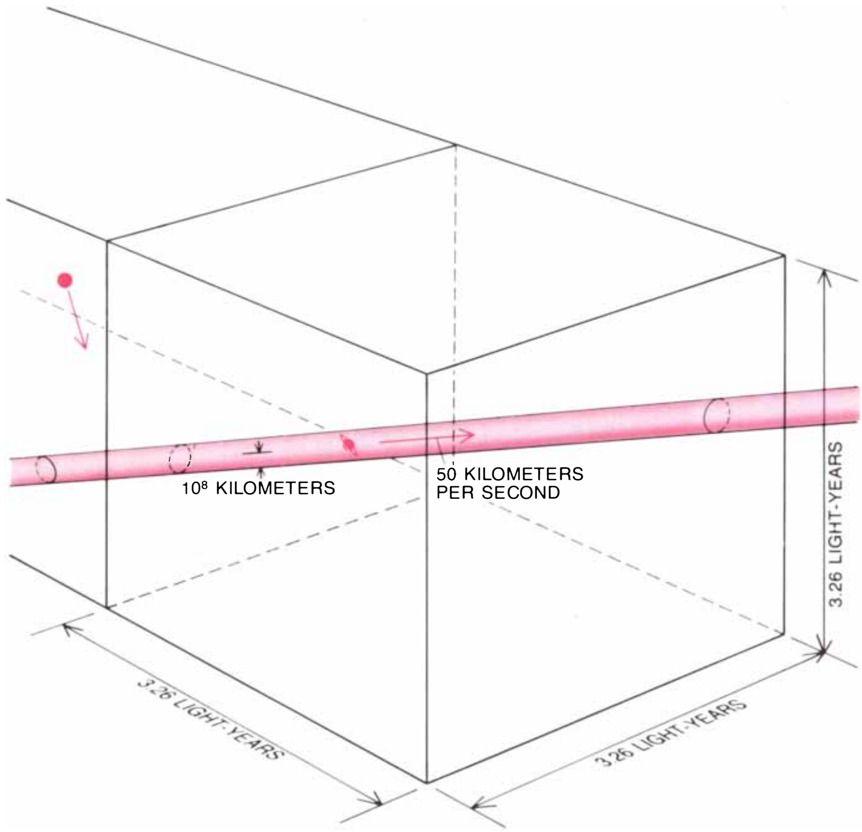
NONLUMINOUS MATTER may constitute a significant part of the total mass of the universe. There is evidence that such matter makes up at least half of the mass of most galaxies. If galaxies were made up entirely of matter that can be observed directly, stars at the periphery of a galaxy would revolve about the galactic center with a velocity that decreases with distance from the center (*left*). The observed velocity of peripheral stars, however, does not decrease with distance. The discrepancy can be explained by assuming that nonluminous matter is distributed throughout the galaxy and in a halo around it. Stars that appear to be on the periphery, therefore, would actually be in the midst of galactic matter, and the distribution of their velocities with distance from the galactic center would more closely match the observed distribution (*right*). The lengths of the arrows in the diagrams are proportional to the velocities of the stars.

The second major event is that all stars will lose their planets. If a star that has a planet is approached by another star to within the radius of the planetary orbit, the orbit will be markedly disrupted by the gravitational field of the passing star, and the planet can be scattered into space. The average time before such an encounter can be expected depends on the density of stars in a given region, on the area of the planetary orbits and on the velocity of the stars with respect to one another. The density of stars can be expressed in terms of the volume in which at least one star is likely to be found. A star moving through space with a planet traces a cylindrical volume whose size depends on the area of the orbit of the planet and on the star's speed. The average interval between stellar encounters is equal to the time required for the volume of the cylinder to become as large as the volume in which a star can be expected [*see illustration on opposite page*].

The density of stars in a galaxy is one star per 35 cubic light-years. Freeman J. Dyson of the Institute for Advanced Study in Princeton suggests that a reasonable value for the radius of a planetary orbit is about 100 million kilometers, and he notes that a typical star pulls its planets through space at about 50 kilometers per second. The volume of the cylinder traced by the moving planetary system will therefore be 35 cubic light-years after about 10^{15} years, and so a disruptive encounter is likely in this interval. It can be safely assumed that after 100 such encounters all of a star's planets will have been kicked out of orbit, so that in about 100 times 10^{15} years, or 10^{17} years, all stars will have lost their planets.

The third transition foreseen is the result of still closer stellar encounters; their effects are evident on a galactic scale. When two stars come close to each other, the gravitational interaction can transfer kinetic energy from one star to the other. If the encounter is close enough, one star can gain so much energy that it attains the velocity needed to escape from the galaxy. Because energy is conserved in the interaction, the second star in the pair must lose kinetic energy, and as a result it becomes more tightly bound to the core of the galaxy.

The process can be called galactic evaporation, because the stellar interactions reproduce on a majestic scale the interactions of molecules evaporating at the surface of a liquid. A similar exchange of energy can also cause an appreciable fraction of the interstellar gas to escape from the galaxy. After as much as 90 percent of the mass in the galaxy has evaporated, the gravitational field will draw the remaining stars and dust into an increasingly dense core. Galaxies in their present form are likely to include a central, supermassive black hole, a region of space from which escape is impossible (except through mechanisms described by the quantum theory). If no such black hole exists now, the galactic core would probably still grow so dense that gravitation would overwhelm the resistance offered by gas pressure, and the core would collapse anyway to form a supermassive black hole. A calculation similar to the one we have outlined for the loss of planets shows that the evaporation of stars and the collapse of galaxies should be complete after about 10^{18} years.



FREQUENCY OF ENCOUNTERS between a star and a planetary system can be estimated from the number of stars per unit volume, the size of the planetary system and its velocity with respect to the stars. The encounters disturb a planetary orbit whenever the star intersects the volume of space filled by the orbit of the planet. On the average such an encounter will take place in the same time required for the planetary system to trace a cylinder in space equal in volume to the cube within which one expects to find a single star. There is one star in the galaxy for every 35 cubic light-years. Each planetary system moves at about 50 kilometers per second, and the outer planetary orbits trace a cylinder whose radius is about 100 million kilometers. Under these assumptions it is straightforward to calculate that one encounter takes place on the order of every 10^{15} years. After about 100 such encounters, which can be expected in 10^{17} years, the orbits of all planets will have been so seriously disrupted that the planets will have escaped from their parent star and scattered into interstellar space. The scale of the stars and the planetary system with respect to their density in space has been greatly exaggerated.

The fourth and fifth transitions we envision for the open universe are late cosmological effects predicted by most grand unified models of particle physics, but the effects do not become important until the universe is 100 times older than it is when galaxies collapse. The aim of the grand unified theories is to present a comprehensive account of the strong, the weak and the electromagnetic forces between elementary particles. At the relatively low energies accessible in terrestrial laboratories the three forces seem quite different; nevertheless, the theories describe them as manifestations of a single interaction whose unity can be perceived at the energy corresponding to a temperature on the order of 10^{27} degrees K. At that extraordinary temperature particles such as the quarks, which "feel" the strong force, can be transformed into particles such as the electron and the positron, which at lower energies feel only the weak and the electromagnetic forces. At lower temperatures such changes of identity become highly improbable, but they can still take place occasionally. The proton, which is thought to be a bound system of three quarks, could decay if its constituent quarks underwent such a transformation.

According to most grand unified theories, all protons should decay after a period of from 10^{30} to 10^{32} years; a decay rate of at least one proton per year should therefore be registered in any aggregate that includes 10^{32} protons, such as a mass of water weighing about 160 tons. There are now 13 experiments either in operation or under consideration that attempt to monitor large masses of water, iron or other materials for evidence of proton decay. Decay signals have tentatively been identified from 150 tons of iron being monitored 2,300 meters underground at the Kolar goldfields near Bangalore in southern India. A single candidate decay signal has also been registered by detectors installed in the Mont Blanc Tunnel between France and Italy. It is too soon to judge whether these signals will continue to stand up to scrutiny.

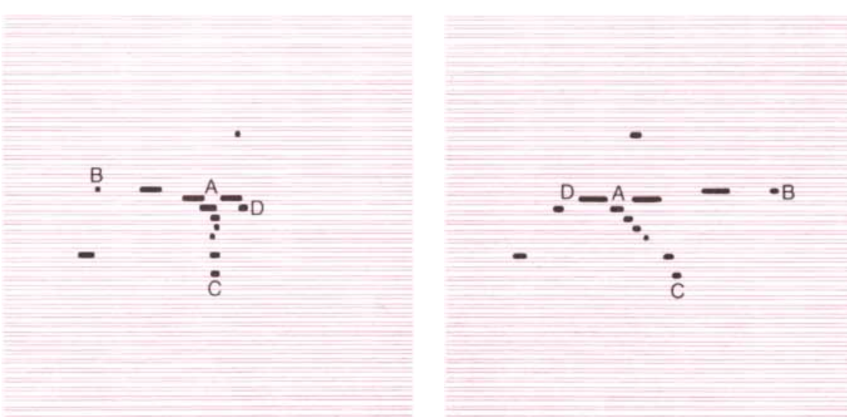
If the proton is susceptible to decay, the process will have a significant effect on stars that have not been captured by galactic black holes. Such stars are the ones that escape by evaporation, and the decay of the protons and neutrons in them will keep them considerably warmer than the surrounding interstellar medium. For an assumed proton lifetime of 10^{30} years the decay rate in a typical star the size of the sun is about 10^{27} per year. Each proton decay gives rise to a shower of energetic electrons, positrons, neutrinos and photons. All the daughter particles except the neutrinos are absorbed by the star, and the absorbed energy keeps the star warm.

The precise temperature of the star during the era of proton decay can be determined by setting the rate at which energy is radiated from the star equal to the rate at which the decay releases heat energy. In this state of equilibrium the temperature depends on the mass of the star, on the surface area from which the heat can be radiated and on the rest energy and lifetime of the proton. The equilibrium temperature is 100 degrees K. for the most massive dead stars (which, paradoxically, are the smallest ones) and about three degrees K. for larger and less massive ones.

The stars will cool to their equilibrium temperature in about 10^{20} years and thereafter their temperature will remain roughly constant until most protons have decayed, in about the year 10^{30}. The stellar glow is cold, but not in comparison with the temperature of the background radiation left over from the big bang. The background temperature depends on the details of the expansion of the open universe. If the density of the universe is less than the critical density, the background-radiation temperature will have fallen to 10^{-20} degree K. by the year 10^{30}. On the other hand, if the density is exactly equal to the critical density, the universe will have expanded more slowly and the temperature of the background radiation will have fallen to 10^{-13} degree K. That is, the temperature will be from 13 to 20 orders of magnitude less than the temperatures of the dead stars.

Proton decay also changes the constitution of interstellar gases that evaporate before galaxies collapse. Inside a star a positron released by the decay of a proton soon encounters an electron, and the two particles annihilate each other. The annihilation generates more photons and heats the star. In intergalactic space the density of matter is so low (and continuously decreasing because of the expansion of the universe) that positrons and electrons are extremely unlikely to collide. Indeed, by the year 10^{30} an open universe with a subcritical density will have expanded to more than 10^{20} times its present size, and the average distance between an electron and a positron in interstellar space will be of the same order of magnitude as the size of our galaxy. (If the density of the universe is equal to the critical density, the universe will have expanded by a factor of 10^{13}.) The interstellar medium therefore becomes an extremely rarefied gas, formed from about 1 percent of today's electrons and from positrons created by the decay of about 1 percent of today's protons.

The events associated with the decay of the proton will have played themselves out by the time the universe has existed for 10^{32} years, or 100 times as long as the average proton lifetime. What remains of the universe then is the rarefied electron-positron gas, the photons and neutrinos that were emitted in various earlier epochs and the supermassive black holes. The photons and neutrinos are left over from the big bang, from the years when the stars shone, from the decay of protons and neutrons throughout history and from the final decay of dead stars. Photons and neutrinos lose energy, and both they and the other constituents become more rarefied as the expansion continues. In other respects the universe at subcritical density remains quiescent until roughly the year 10^{100}, a period lasting 10^{68} times as long as all the processes we have described so far.

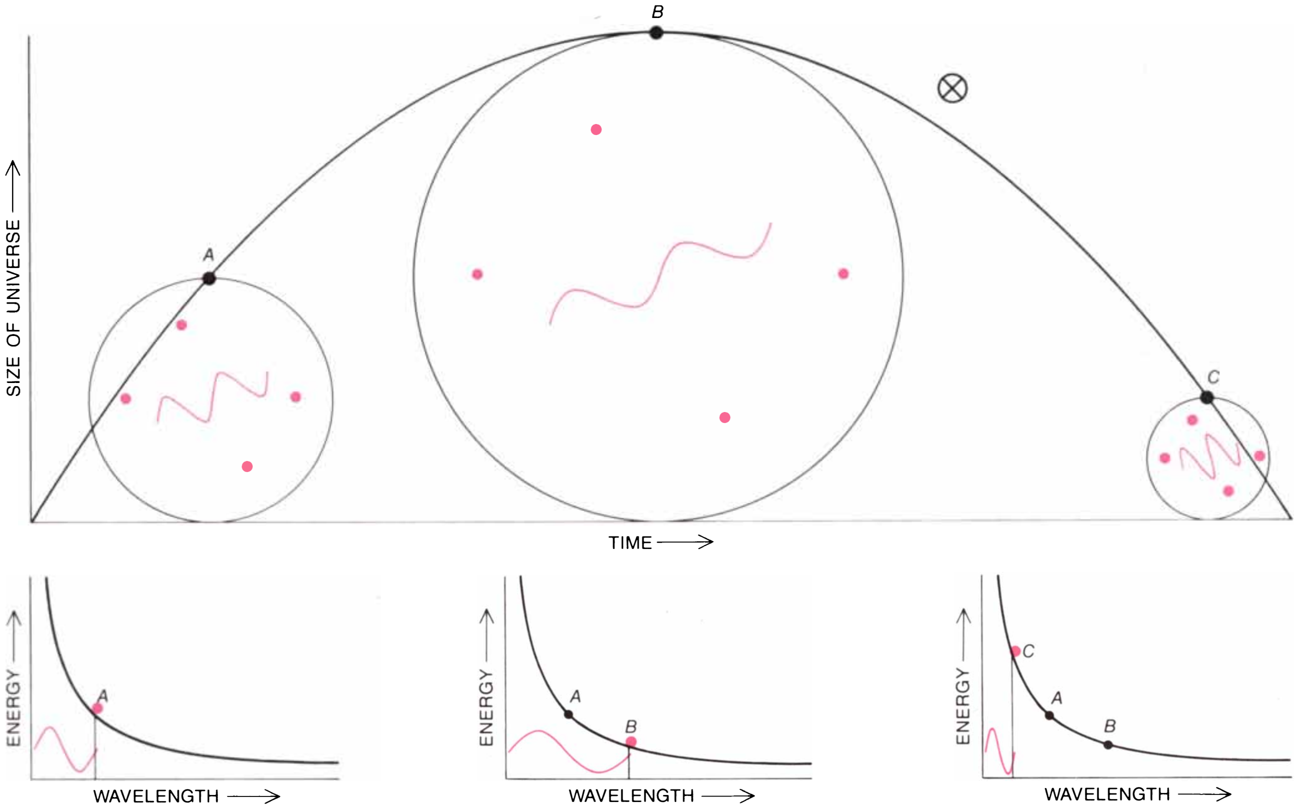


DECAY OF THE PROTON is a possible event that would have important consequences for the future of the universe. The decay should be observed experimentally if it is not too rare. Most of the current "grand unified theories" of elementary-particle interactions predict that the lifetime of the proton is between 10^{30} and 10^{32} years. Some 13 experiments now under way or in the planning stages are trying to detect the decay. Candidate decay events from two experiments are under study. Some of the events were detected by investigators working in a mine in the Kolar goldfields near Bangalore in southern India; another event was seen by investigators from the European Organization for Nuclear Research (CERN), working with a detector in the Mont Blanc Tunnel between France and Italy. The Mont Blanc event is shown in two schematic views at a right angle to each other. The parallel planes are plates of iron that are monitored for decays, and the black dots represent the products of decay. If proton decay can be verified, an evolutionary milestone can be fixed at about the year 10^{30}. All long-lived matter except electrons, positrons, photons, neutrinos and black holes will have decayed by then.

The sixth and last major event in the future of the open universe is the decay of the black holes. In the simplest interpretation of Einstein's gravitational theory nothing can escape from a black hole. There is a boundary called the event horizon where the velocity needed to escape is equal to the speed of light; no particle inside the event horizon can ever go fast enough to cross it. Nevertheless, in 1974 S. W. Hawking of the University of Cambridge showed that a quantum-mechanical phenomenon implies that a black hole can give up all the energy associated with its mass and thereby disappear.

The quantum phenomenon is a consequence of the uncertainty principle of Werner Heisenberg, which states that it is only with a certain probability that either the momentum or the position of any particle can be determined to be within a given range of values. More precisely, the product of the uncertainty in the momentum of the particle and the uncertainty in its position is not less than a numerical constant. In classical physics (that is, in physical theory that does not take account of quantum phenomena) a particle can cross an energy barrier only if it is given enough energy to surmount the barrier. The event horizon is an absolute energy barrier in classical physics: a particle could not gain enough energy to surmount it. Because of the uncertainty principle, however, a particle initially found in one region can later be found in another region, even if its classical energy is much less than the height of the energy barrier between the two regions. A particle that crosses an energy barrier in this way, without acquiring the energy necessary to surmount it, is said to have tunneled through the barrier. Hawking pointed out that because particles can tunnel through the event horizon, mass and energy can be ejected from a black hole [see "The Quantum Mechanics of Black Holes," by S. W. Hawking; SCIENTIFIC AMERICAN, January, 1977].

Hawking showed that the rate at which energy is emitted by the black hole is inversely proportional to the square of its mass. The rate is initially low, but as the mass decreases, the energy loss accelerates. It follows that all black holes must eventually disappear, or evaporate; for the supermassive black-hole relics of collapsed galaxies the evaporation time is about 10^{100} years. Most of the decay products are photons; the emissions become increasingly energetic during the later stages of the evaporation. After 10^{100} years, therefore, the universe is made up of an extremely diffuse gas of electrons, positrons and neutrinos, low-energy photons emitted long before black-hole decay and numerous expanding spheres of



ENERGIES OF FREE PARTICLES are not conserved in a closed universe. Because of cosmic expansion and contraction in such a universe, phenomena that depend on distance vary with time, just as the distance between two points on the surface of a balloon increases or decreases as the balloon is inflated or deflated. Thus the wavelength of a photon increases during expansion and decreases during contraction. The energy of the photon is proportional to the reciprocal of the wavelength, and so the energy of the emitted photon decreases during the expansion and increases during the contraction. When the wavelength of the photon is shorter than it was at the time of its emission, there is a net gain in the energy of the universe as a whole. The additional energy could prolong a new, subsequent expansion.

high-energy photons emitted by evaporating black holes.

Mathematical models of the large-scale structure of an open universe suggest that the universe we live in is a very special place. To describe the universe mathematically six quantities must be specified: three of them give the rate of expansion along the three directions in space and three give the rate of change in the expansion rate (that is, its acceleration or deceleration) along the three directions. If the six quantities are specified at some definite time and if the distribution of energy is given for the same time, the values of the six quantities and the distribution of energy are determined for all later times.

In 1973 Hawking and C. B. Collins, also of Cambridge, showed that for almost all possible values of these quantities at an early stage in the big bang the universe must become increasingly anisotropic as it expands. The average separation of galaxies along one spatial axis ought to become increasingly different from the average separation along the other axes because of differences in the expansion rates. The result suggests that the actual universe is a highly improbable one. For example, as we noted above, the cosmic microwave background is invariant with direction to within one part in 10,000. On a cosmic scale the matter in the universe is also smoothly distributed. A major unsolved problem in cosmology is to understand why the universe is so nearly isotropic.

One approach to the problem is to choose initial values of the six quantities in such a way that the energy density of matter and radiation is exactly equal to the critical density. If such a universe is dominated by matter instead of radiation throughout its history, any initial anisotropy will not grow. The decay of the proton, however, can lead to a radiation-dominated universe.

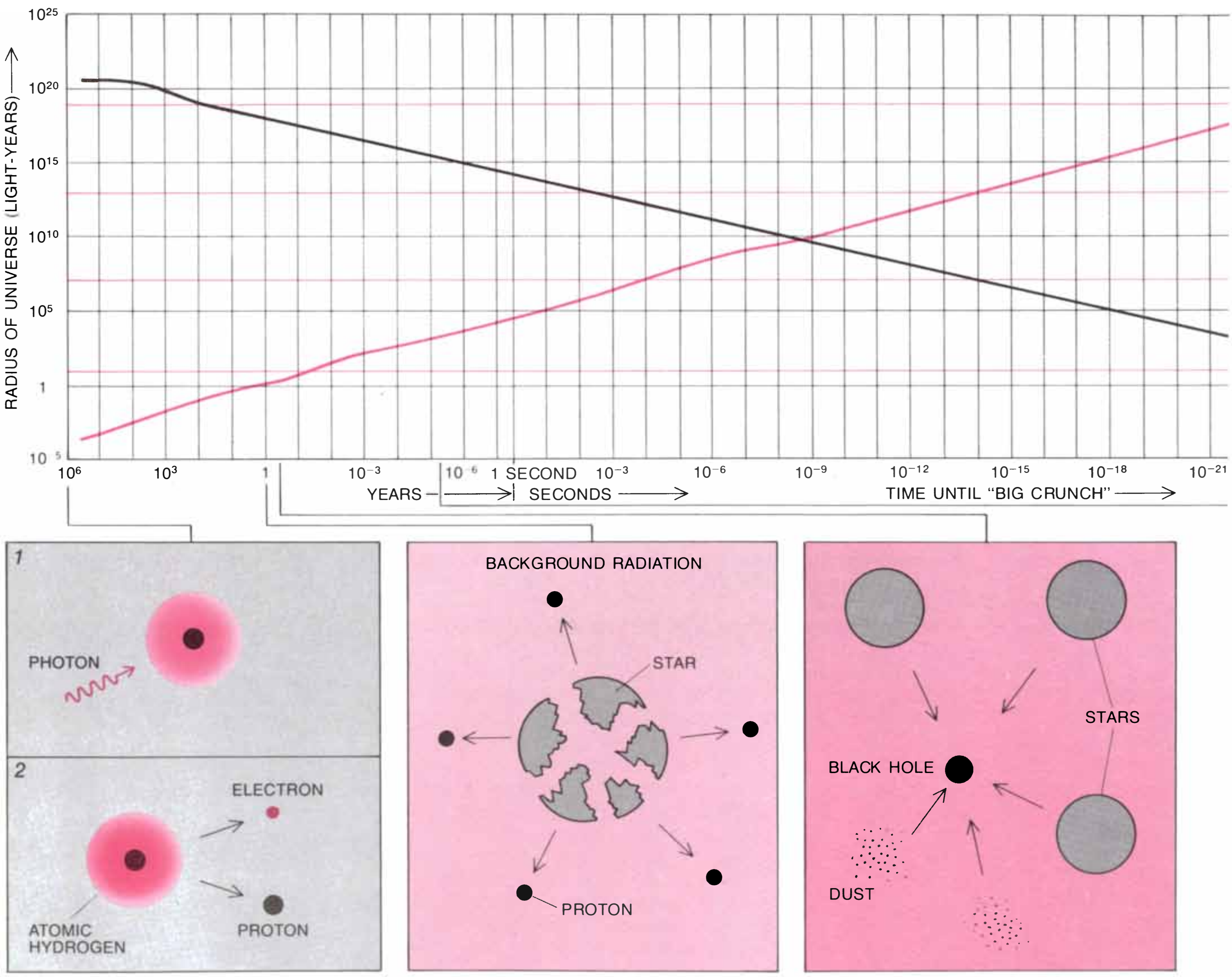
John D. Barrow and Frank J. Tipler of the University of California at Berkeley have shown that in a universe at the critical density the electrons and positrons remaining after proton decay will begin to form bound pairs after more than 10^{70} years. Each bound pair is effectively an atom, but the electron and positron orbit each other in such a highly excited state that the region of space encompassed by the pair is larger than the present observable universe. The encompassed space grows progressively larger the later a bound pair is formed.

As time goes on the electron and the positron spiral inward and eventually annihilate each other, giving rise to high-energy photons. The wavelength of the photons is shifted toward the red end of the spectrum as the universe continues to expand. In the model described by Barrow and Tipler the universe finally comes to be dominated by radiation and the anisotropy increases. In a subsequent investigation, however, Don N. Page and M. Randall McKee of Pennsylvania State University showed that under certain conditions the electron-positron pairs can continue to dominate the energy of the universe even though their number is decreasing. The red shift of the photon radiation decreases its energy and the increasingly long time required for the annihilation of late-forming bound pairs causes a decrease in the rate of photon emission. Hence the anisotropy that is characteristic of a radiation-dominated universe may not be able to grow.

All the foregoing speculations concern an open universe. We shall now consider the future of the universe on the assumption that there is enough nonluminous matter for gravity to halt the expansion and bring about a contraction. The closer the average density is to the critical density, the longer the closed universe will last. We know of no reason, however, for the average density to be close enough to the critical density for the universe to survive until most protons decay. The universe at its maximum expansion is therefore likely to be made up of dead stars, supermassive black-hole remnants of collapsed galaxies and low-energy photons and neutrinos, just as it would be if the universe were open.

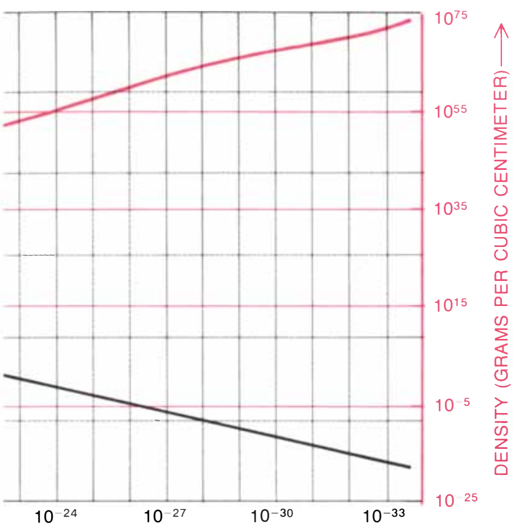
One curious aspect of a closed universe is that although energy is conserved locally, the total mass or energy of the universe is not conserved. For a given size of the universe the total energy during contraction is greater than it is during expansion. Consider a photon emitted from the sun toward intergalactic space. The conservation law is satisfied during the emission because the energy carried by the photon is exactly balanced by a very small decrease in the mass of the sun. As the universe expands, the wavelength of the photon increases in proportion, with the result that its energy diminishes. When the universe contracts, the photon's wavelength contracts as well and its energy increases. Eventually the wavelength of the photon becomes shorter than it was at the time of emission, so that the photon has gained energy without any compensatory loss of mass or energy in some other region. The universe therefore runs hotter in contraction than it does in expansion. The photons that make the largest contribution to the extra energy are the ones emitted when the universe is at or near its maximum size, because such photons will undergo the most contraction.

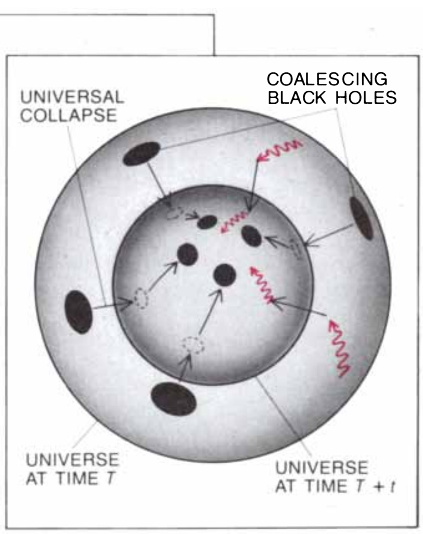
The major events during the expansion phase of a closed universe follow the same sequence as the events in an open universe. (There is, of course, no collapse for an open universe.) Several investigators have studied the collapse, including Martin J. Rees of Cambridge. As the photons gain energy during the contraction they heat the dead stars, causing them to begin to burn rapidly, explode or evaporate. The resulting



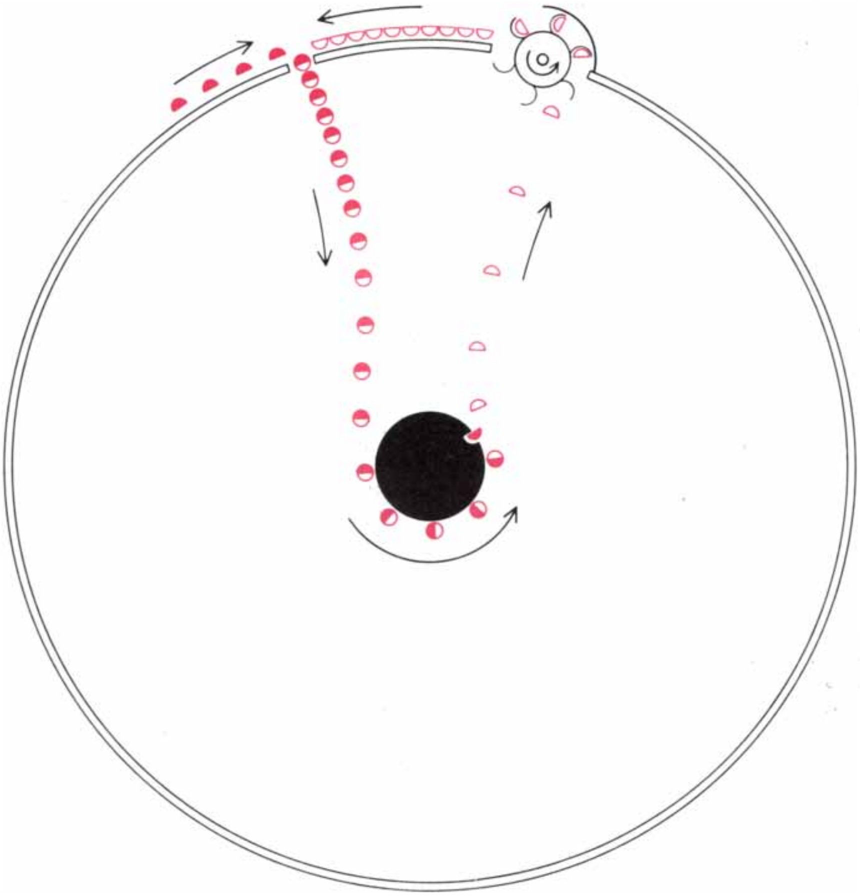
FINAL STAGES in the collapse of a closed universe would very roughly reverse the stages in its expansion if it were not for the effects of black holes. About 20 billion years before the big crunch, or the complete gravitational collapse of a closed universe, the universe will contract to a volume at which its energy density will be the same as it is now. As the contraction continues photons will become more energetic and the universe will heat up; about a million years before the big crunch the photons will dissociate interstellar hydrogen atoms into their constituent electrons and protons. A year before the crunch the temperature outside stars will become greater than the temperature inside, and the stars will begin to break up. At about the same time supermassive black holes will begin to swallow up the remaining stellar material as well as other matter and radiation. About three minutes before the crunch the supermassive black holes will be-soup of particles would continue to retrace the steps in the expansion of the universe if it were not for the effects of the black holes. As the density increases, the black holes gobble up material, and they coalesce whenever they collide. One can calculate that in a universe with one supermassive black hole per galaxy the dead stars are swallowed by black holes shortly after the stars break up and material begins to evaporate from them. All the black holes finally coalesce into one large black hole that is coextensive with the universe.

Theoretical physics in its current state of development cannot fully describe the collapse of a black hole; it is not yet known how to extrapolate the

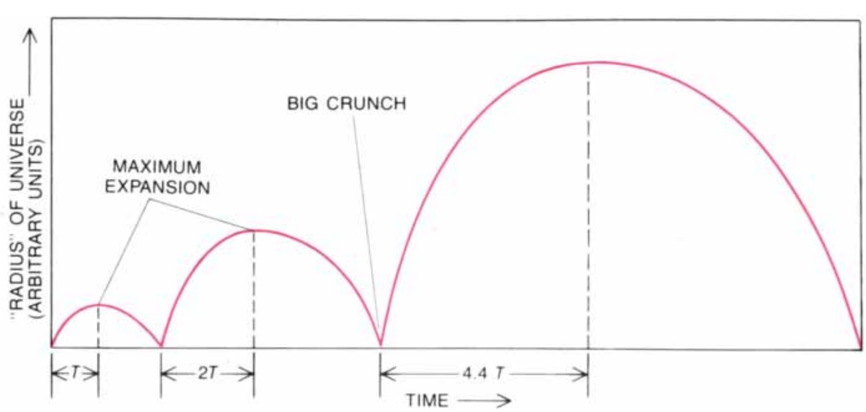




gin to coalesce; the coalescence is represented on the contracting surface of a sphere, which corresponds to the contraction of space. The variation in the density and radius of the universe is graphed with time; it was calculated numerically by applying Einstein's equations and other principles described in the text.



COSMIC ENERGY CRISIS could be delayed for a long time by the existence of black holes formed from collapsing galaxies. In principle energy can be extracted from the gravitational field of a black hole; hence it is conceivable that an advanced civilization could maintain itself for at least 10^{100} years, or as long as it takes for black holes to evaporate. The most efficient known scheme for extracting the energy of a black hole was put forward by Charles W. Misner of the University of Maryland at College Park, Kip S. Thorne of the California Institute of Technology and John Archibald Wheeler of the University of Texas at Austin. A rigid shell is constructed around a rotating black hole, and surplus material from the civilization is sent by a satellite close to the event horizon, the boundary of the region from which nothing can escape. Near the event horizon the material is jettisoned at the equator of the black hole. As a result both the gravitational potential energy of the jettisoned material and the rotational energy of the black hole are decreased. Because the total energy of the satellite and the black hole must be conserved, the energy of the satellite is increased by an amount equal to the two energy reductions. In principle the energy gained by the satellite could be converted into a useful form.



UNIVERSE MAY BE CYCLIC if the density of matter is sufficient to close it and if physical processes as yet unknown allow the universe to "bounce," or expand once again. Because energy is gained during the contraction phase, each new cycle expands for a longer time than the preceding one. The period of expansion should increase by a factor of about two during the next cycle of the universe, assuming that the current expansion continues for about 100 times the present age of the universe. The duration of succeeding cycles increases by larger factors.

equations that govern the large-scale structure of the universe back to a state of infinite density. It is possible, however, that before the density becomes infinite an unknown mechanism could cause the universe to "bounce" and begin expanding once again.

If the universe can bounce, it could remain closed and so be cyclic. The energy gained by the photons during each period of contraction might be conserved through the bounce; with each successive cycle the universe would therefore be larger, for a given temperature, than it was in the previous cycle and would also take longer to reach its maximum size. Robert H. Dicke and P. J. E. Peebles of Princeton University and the four of us have calculated that, if our universe is cyclic, the next cycle should expand for about twice as long as the expansion phase of the current cycle. For earlier cycles the expansion factor would be smaller; we calculate that the present universe is at most 100 bounces from the cycle that lasted just long enough to make a single generation of stars. The major difficulty for the bounce model is to understand how an extraordinarily inhomogeneous and locally anisotropic universe of isolated, coalescing black holes can be smoothed out; for that matter, there is no understanding of how a bounce can take place.

If the universe is cyclic, and if reactions mediated by the high temperatures available at the beginning of each cycle give rise to the excess of nuclear particles over antiparticles that is observed today, one can draw a strong conclusion. The evolution of the universe from the beginning of each bounce until the formation and dissolution of galaxies will follow the same course of events during each cycle. As the cycles become longer, however, the phenomena we have described for the later stages of the open universe will become important. For example, on the first cycle in which the expansion stage lasts longer than a ten-billionth of the average proton lifetime (that is, longer than about 10^{20} years) the energy gained during the contraction stage will be dominated by the photons emitted in the course of proton decay. Subsequent cycles may then lengthen by a factor of 1,000 instead of a factor of about two. Even larger expansion factors, on the order of 10^{12}, might result from the effects of coalescing black holes.

From a human point of view perhaps the most important question one can ask about the universe concerns the future of life and intelligence. Can the habit of thinking persist indefinitely in a universe increasingly hostile to life as it exists on the earth? Several cosmologists, including Dyson and Steven C. Frautschi of the California Institute of Technology, have begun to speculate about issues such as the energy requirements for the indefinite maintenance of life and for communication among members of a society dispersed across increasingly large regions of space.

Dyson assumes that the embodiment of life and consciousness need not be limited to cells and DNA; instead, he maintains, the essential feature of consciousness is complexity of structure, which can be realized in any convenient material. Thus he assumes that the idea of a sentient computer or a sentient cloud cannot simply be dismissed as philosophically incoherent.

Given these assumptions, the changes in the environment caused by the death and cooling of stars and their evaporation from galaxies need not be insurmountable to a system that could be counted as alive and intelligent. For example, in principle energy could be extracted from the gravitational field of a supermassive black hole. The decay of protons and neutrons, however, may bring about a fundamental change, since it seems unlikely (although perhaps not impossible) that intelligence could be based on the tenuous foundation of electrons and positrons. Moreover, if the universe is closed, the conditions necessary for life may exist only for certain periods during each cycle.

In an open universe the ultimate impediment to life is quite different. With the evaporation of black holes there will be a cosmic energy shortage, because as the expansion of the universe continues the remaining particles and photons lose energy. Any constant rate of energy consumption by life forms will in due course become untenable. On the other hand, Dyson proposes that increasingly long periods of hibernation, during which energy is not consumed, could be interspersed with periods of consumption. Hence there is the potential for extraordinarily long-lived civilizations in this universe.

The starting point of contemporary cosmology, as we have emphasized, is the big-bang theory. Our quantitative treatment of the evolution of the big bang into the remote future is based on the new theories that describe the physics of elementary particles. In spirit, however, these ideas were expressed by the 11th-century Persian astronomer and poet Omar Khayyám, whose poem *The Rubáiyát* was translated by the English writer Edward FitzGerald. In one quatrain Omar wrote:

With Earth's first Clay They did
  the Last Man's knead,
And then of the Last Harvest sow'd
  the Seed:
Yea, the first Morning of Creation
  wrote
What the Last Dawn of Reckoning
  shall read.

Memory in Food-hoarding Birds

Birds that hide seeds and later recover them appear to excel in spatial memory. One species evidently remembers where it has put thousands of caches for as long as several months

by Sara J. Shettleworth

In England in the winter a bird feeder stocked with peanuts is all the apparatus one needs to get an animal to show behavior suggesting it has remarkable powers of memory. Indeed, its memory appears to far surpass in capacity the memory that any other animal has ever shown in the laboratory. The feeder attracts many birds. Among them are great tits, blue tits and marsh tits, which are all small, lively birds related to the North American chickadees. The great tits and the blue tits congregate at the feeder, eating as fast as they can. They interrupt their meal only to chase away their competitors. A marsh tit nonetheless darts in, grabs a peanut and flies off. It is back almost immediately to grab another. It stores the peanuts nearby, each in a different site, until the feeder is empty. Then it searches out its hidden food.

In the American Southwest another bird, Clark's nutcracker, a relative of the jays and crows, shows similar behavior. In the late summer the nutcracker harvests the seeds of piñon pines. It repeatedly fills its sublingual pouch, a pocket under the tongue, and then flies several miles to bury the seeds. Often the burial sites are on bare, south-facing slopes where the snow will not be deep later on. A nutcracker may bury as many as 33,000 piñon-pine seeds in caches of four or five seeds each. Throughout the winter it returns and digs up its thousands of caches.

How do these birds find their hoards? Does the tit or the nutcracker remember where it has stored each peanut or cache of piñon seeds? Until recently observers of hoarding behavior in the wild doubted that the birds rely on memory. For one thing the bird would need a capacious memory indeed to remember the sites of hundreds or thousands of individually hoarded items. The memory would also have to be long-lasting. Even a short-term hoarder such as the marsh tit does not recover its stores until hours or days after it deposits them, and a long-term hoarder such as Clark's nutcracker does not return to a hoard for months, perhaps not until spring, when nutcrackers feed their young on piñon seeds. Moreover, in principle the hoards could be recovered without the aid of memory. A hoarder could store food only in certain kinds of sites, such as south-facing slopes or holes in bark. To recover its stores it would need only to search in sites of that kind. Then too it could employ memory only for the area in which it hoarded rather than for the individual sites. Inside that area it could search by trial and error or conceivably by cues such as odor.

Psychologists studying learning and memory in animals are becoming increasingly aware, however, that certain species have adaptive specializations that make them particularly good at learning and remembering things it is important for them to know. Among the well-known instances are the ability of many birds to learn their species' songs, the ability of rats and other animals to remember spatial locations, the ability of bees to remember the location of flowers and the ability of many animals to learn to avoid noxious food. If food-storing birds really do remember large numbers of storage sites over long periods, their memory could be another example of an adaptive specialization, one that would enable the birds to recover their stores efficiently. A bird that could remember where it had hidden its food would make fewer errors recovering it and spend less time and energy than birds searching randomly. Recent studies indicate that at least some food-storing birds do remember the sites of their hoards quite well.

Some of the most detailed evidence that food-storing birds remember their hoards has been accumulated over the past few years by John R. Krebs and his associates at the University of Oxford. They began by looking for evidence confirming that marsh tits in their natural environment recover their hoards. To that end marsh tits were trained to come to sunflower-seed dispensers set up in their territories in Wytham Wood near Oxford. Then Richard J. Cowie, Krebs and David F. Sherry coated the husks of the seeds with a radioactive substance harmless to the birds. (The birds remove the husks before eating the seeds.) The radioactivity enabled the investigators, equipped with a portable scintillation counter, to find the seeds the birds had cached and then check every few hours to see when they disappeared.

The seeds of course might be eaten by other birds or by rodents such as mice and voles. Cowie, Krebs and Sherry reasoned, however, that if marsh tits use their memory to recover their own stores, the seeds should disappear sooner from the birds' storage sites than they would from false sites the investigators established. This was indeed the case. The seeds the birds stored were depleted sooner than seeds the investigators placed in similar sites that were each one meter from a hoard made by a bird. In Wytham Wood in the winter the natural hoards were generally gone in a day or two. The last of the false hoards disappeared fairly soon afterward, suggesting that predation of the hoards is common.

One defensive tactic the marsh tits apparently adopt against animals that steal from their hoards is to disperse seeds rather than burying them in clusters. On the average the hoards the marsh tits established in Wytham Wood were seven meters apart. Sherry, Mark Avery and Allen Stevens set up false hoards of sunflower seeds and found that the seeds stayed in place longer the farther apart they were, up to a distance of about seven meters. Hoards farther apart than that were no safer than the ones at seven meters. In Wytham Wood, at least, marsh tits seem to have hit on the optimal spacing.

Having obtained evidence that marsh tits in nature recover their hoards, Krebs and his colleagues turned to observations in the laboratory. Marsh tits supplied with a bowl of seeds and with suitable storage sites hoard quite readily in captivity. Sherry, Krebs and Cowie offered marsh tits trays of moss one meter on a side in which to store sunflower seeds. The birds did appear to remember where they had buried seeds even after 24 hours. In these experiments, however, the behavior of the birds was recorded only as visits to relatively large areas on the tray. Moreover, in an effort to ensure that the birds employed only memory in the experiments, the hoarded seeds themselves were removed from the tray before each bird was allowed to return. Thus it was difficult to say with what precision the birds located their hoarding sites. It was equally difficult to identify mistakes the birds might be making.

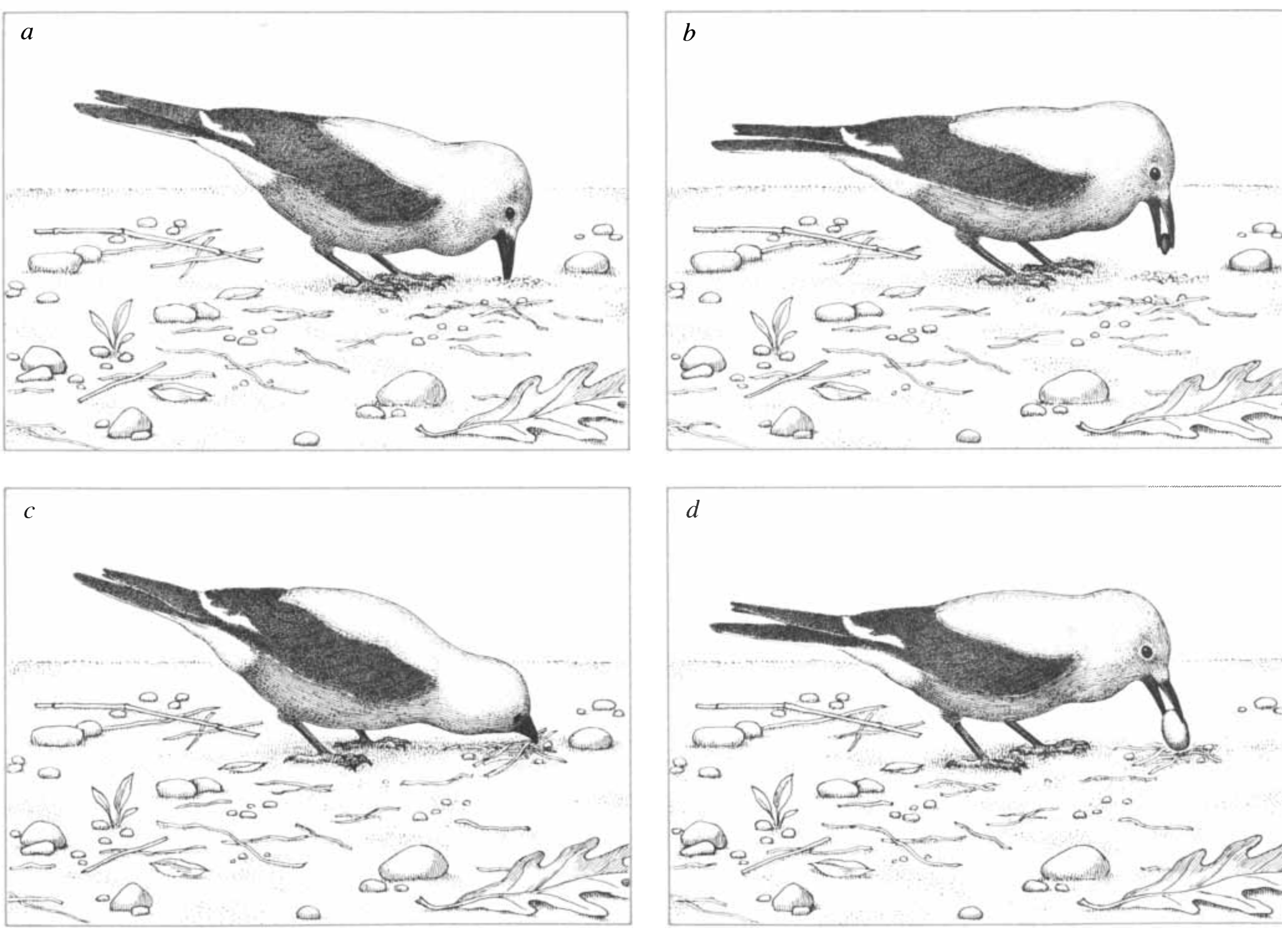
When a Guggenheim Fellowship enabled me to spend a sabbatical leave at Oxford, my interest in adaptive specializations of learning and memory attracted me to the marsh-tit studies. Krebs and I designed some experiments to analyze in more detail the marsh tits' apparent memory. We were particularly interested in finding out whether the items most recently stored by the birds are the ones they recover first. Such a "recency effect" often appears in tests of memory, and in the case of food hoarding there is good reason to expect it: if hoards are lost to scavengers at a constant rate, the most recent hoards will be the ones most likely to be available to the hoarders at any given time. Looking for them first is therefore the hoarders' best bet.

In a large room Krebs and I set up sections of tree branches. In each section we drilled a number of holes the right size for storing hemp seeds. There were about 100 holes in all. Each hole was covered with a flap of cloth, which a bird would have to lift in order to store a seed or look for one.

The room was the site of a series of experiments. For each trial of our first experiment we allowed a marsh tit to store 12 seeds it got from a bowl on the floor in the middle of the room. Each hole could hold only one seed; thus each seed was stored in a different hole. Once the seeds were cached we kept the bird outside the room for about two and a half hours. We removed the bowl of seeds; then we readmitted the bird, and it searched for the seeds it had stored. If it searched at random among the 100 holes, it would have to investigate an average of about eight holes to find one seed. The birds were much more efficient. On the average each bird made about two errors per seed. At the beginning of a recovery test a bird sometimes went to three or four seeds in succession without looking in empty holes.

Maybe instead of employing memory the tits were smelling the seeds or detecting them in some way we had not thought of. Smell seemed unlikely; the sense of smell is generally poor in birds. Still, to test these possibilities we allowed a marsh tit to store 12 seeds and then we moved the seeds to other holes (ones the bird itself had used in earlier trials). Under these circumstances the readmitted birds made about six errors per seed. They eventually found about two-thirds of the seeds by looking in more holes than usual.

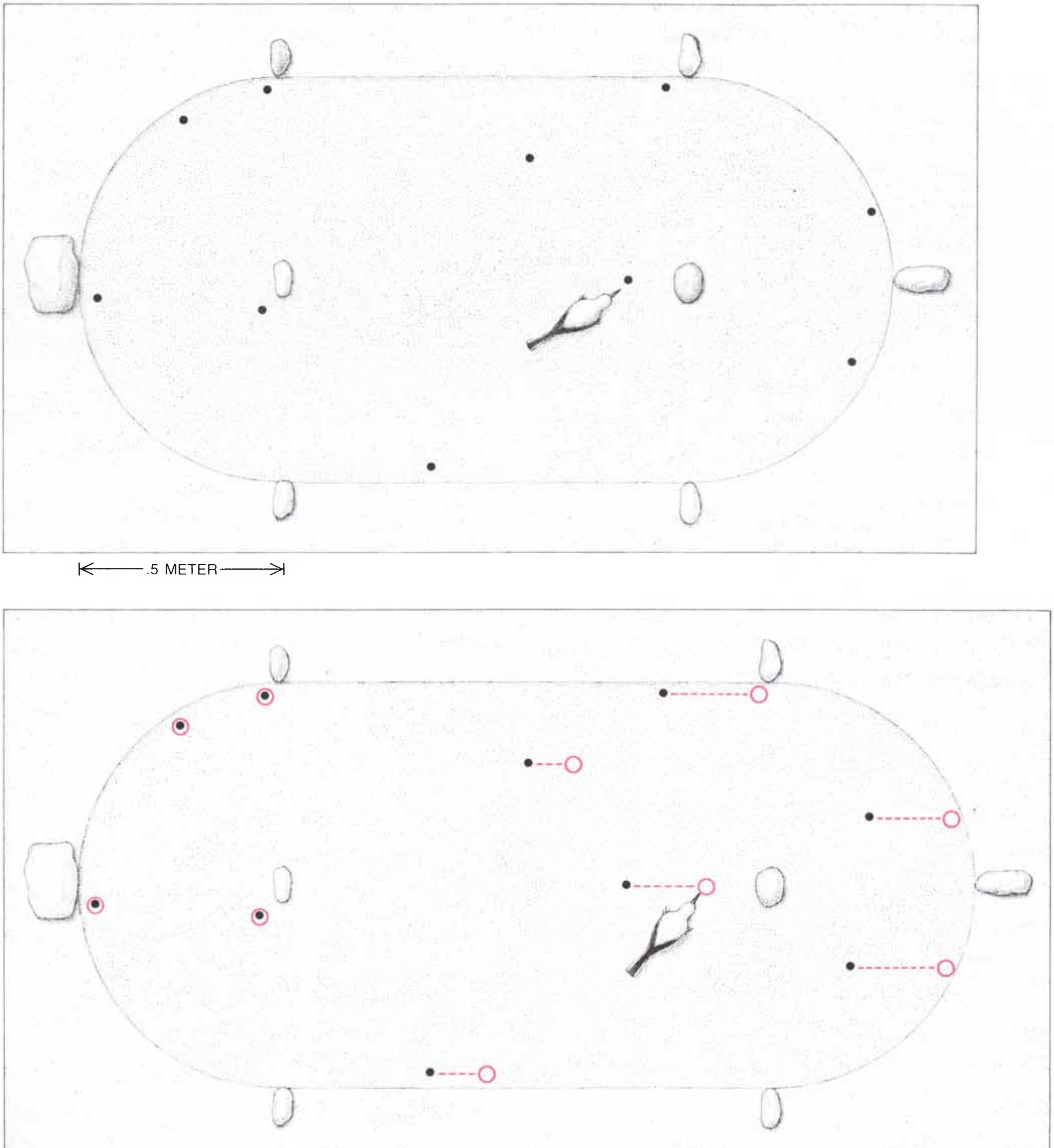
The results indicate that the marsh tits did not detect the stored seeds by smell



HOARDING OF FOOD by a Clark's nutcracker is shown in a sequence of drawings based on observations made by Stephen B. Vander Wall at Utah State University. The nutcracker jabs its bill into the ground to loosen the soil (*a*); then it inserts a conifer seed (*b*), rakes soil and grass over the seed (*c*) and places a pebble on top of the cache (*d*). The pebble apparently serves only to camouflage the site until weather obliterates all traces of the caching. The procedure takes from 10 to 20 seconds. In nature the nutcracker hoards as many as 33,000 piñon-pine seeds in thousands of widely distributed caches that average four or five seeds each. The bird does its hoarding from September through November and relies on memory to find the caches throughout the following winter, spring and summer.

or some similar cue. On the other hand, the performance of the tits in the first experiment does not necessarily mean that they remember storing each seed in a certain hole out of the 100. To some extent each bird in the laboratory grew to have favorite holes in which it stored seeds particularly often. Each bird was likely to inspect its favorite holes when it was searching for stored seeds. The recovery of seeds from such holes could be due to nothing more than the bird's habit of visiting those places both when it stores seeds and when it recovers them. Indeed, this tendency probably explains why the birds' performance was better than chance when we moved the stored seeds.

A closer look at our data showed that a tit did not simply go to the same holes on each test. Even the holes the tit pre-



PROOF OF MEMORY in Clark's nutcrackers emerged from experiments done by Vander Wall. In the experiment shown here a nutcracker was allowed to cache seeds (*black dots*) within an oval area on the sandy floor of an aviary; four stones were arrayed at each end of the oval (*top drawing*). The oval was then lengthened by 20 centimeters and the stones at the right were moved accordingly (*bottom drawing*). Two days later the bird was allowed to search for its caches. The bird dug with success for the caches at the left (*colored circles*). At the right, however, it made errors of approximately 20 centimeters. Evidently it was relying on its memory of the positions of nearby stones. Errors in the middle of the oval suggest the bird estimated the position of some caches with reference to both sets of landmarks.

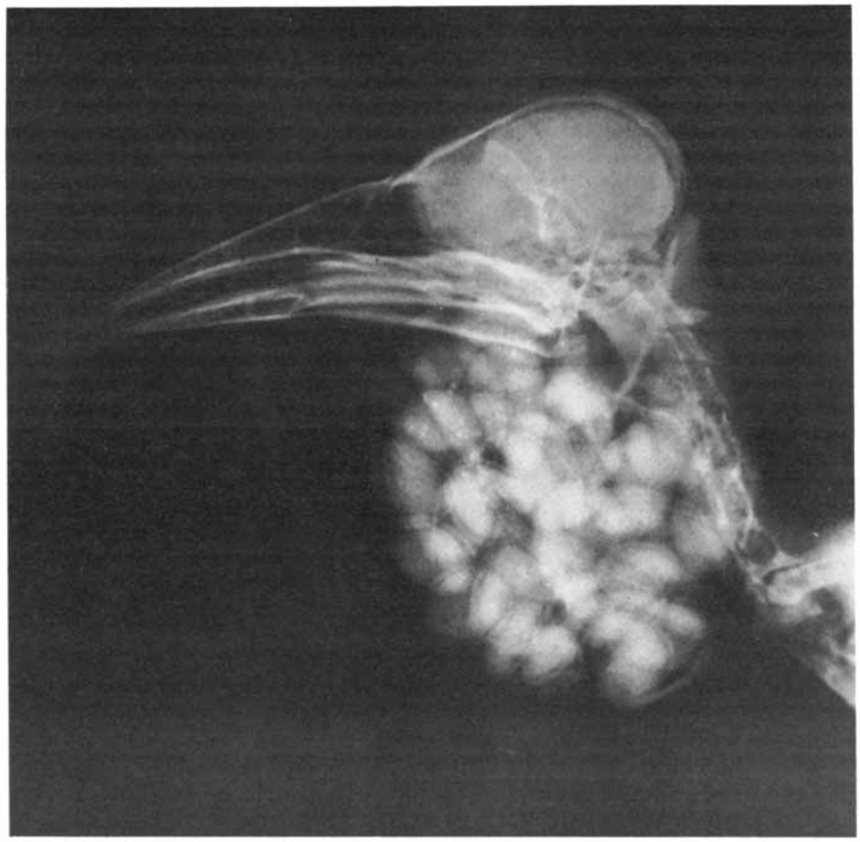
ferred most were more likely to be inspected during the recovery of seeds if the tit had actually put a seed there. Nevertheless, the problem was fundamental. In conventional laboratory tests of memory the experimenter supplies the information the animal must try to remember. In our tests the birds themselves produced the information when they stored seeds where they chose. Any tendency they might have to look into the same holes on successive visits to the test room would lead them to perform as if their memory for the storage sites were good.

Our next experiment was therefore designed to make the birds' memory work against their tendency to prefer certain storage sites. Again each bird stored seeds, was removed from the room and then readmitted some two hours later. Instead of requiring the bird to search for the seeds during the second session, however, we now allowed it to store additional seeds. To get it to do so we simply left a bowl full of seeds in the room. We reasoned that if the tits remembered which holes already had seeds, they would avoid those holes when they stored a second batch. Conversely, if they inspected the same holes each time they visited the room, many of the holes they inspected on their second visit would already have a seed in them.

The results were clear. The birds almost never inspected holes containing a seed when they were storing a second batch of seeds. When they were hungry, however, and there were no more seeds to store, they showed their usual efficiency in recovering seeds from holes. We were interested to note that a bird storing a second batch of seeds often chose a hole near one it had employed in storing the first batch. Thus the bird seemed to remember and avoid the individual holes, not just particular branches or parts of the room.

After a bird had stored two batches of seeds, we could let it recover all the seeds. In this way we could test for the recency effect. If marsh tits show the effect, they should recover seeds from the second batch sooner or more often than seeds from the first batch. Our results suggest that there is a recency effect but that it is not strong. If the memory of storage sites fades slowly, however, the two episodes of hoarding would have to be separated by more than two hours for the recency effect to be clear.

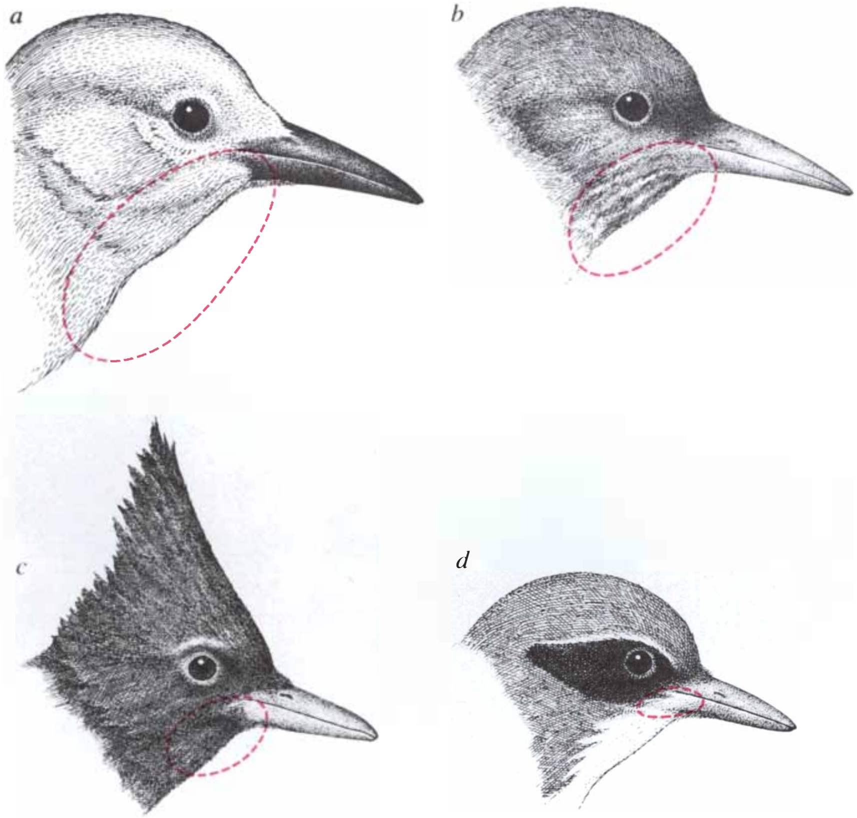
An additional finding emerged from the experiments. It is plain that an efficient searcher for stored food should seldom inspect empty sites. Furthermore, it should be able to remember where it has already been to recover food so that it can avoid going back. Laboratory rats faced with the task of collecting food from a number of sites in a large maze prove to have the latter ability. Marsh tits have it too. In the experiments that Krebs and I did, the holes a bird inspected as it recovered seeds were inspected a second time much less than one would expect if the bird were not remembering and avoiding such sites. In experiments done by Sherry, birds tended not to search areas of moss they had already searched for seeds. A marsh tit searching for stored food makes use, then, of two kinds of information: it remembers where food has been stored and also the sites it has already inspected.



SUBLINGUAL POUCH of a Clark's nutcracker is a pocket under the tongue in which the bird carries seeds. It is thus an anatomical specialization that makes the nutcracker well suited for its life of transporting and hoarding seeds. When Vander Wall made this X-ray photograph, the pouch contained 38 seeds with a net weight of 30.6 grams, or more than an ounce.

If information about hoarding sites is stored in the marsh tit's brain, the bird's memory should be subject to a peculiarity of the nervous system that seems to be present in many birds. In many vertebrate animals (including man) information about the left half of the visual field comes from both eyes, but it goes to the right half of the brain, and vice versa. The situation is somewhat different in birds such as the marsh tit, where the eyes are on the sides of the head and each eye surveys a separate visual field. The tracing of information pathways in the brain of the bird and the results of experiments done mostly with pigeons suggest that in these birds information reaching the brain by way of one eye is stored primarily in the half of the brain on the opposite side of the head. In short, such birds have little or no interocular transfer for many kinds of information. Sherry, working with Krebs and Cowie, employed this fact to interpret an intriguing behavior marsh tits show when they store a seed. Having tapped or poked the seed into place, the bird quickly cocks its head first to one side and then to the other. It is as if the bird were looking at the storage site or at the landmarks near the site with each eye in turn.

Does the marsh tit cock its head in order to store visual information about hoarding sites in both halves of the brain? To answer this question Sherry had marsh tits store sunflower seeds wherever they chose in a tray of moss while each bird was wearing a translucent cover over one eye. If the bird was later allowed to search for the seeds with the same eye covered, it found the seeds quite normally. If the patch was moved to the other eye, however, so that the bird had to search with the eye that had not viewed the storage sites, the bird behaved as if it did not remember where the seeds were. If marsh tits do have



ANATOMICAL SPECIALIZATION of four birds in the southwestern U.S. corresponds to their varying degree of dependence on cached food. Clark's nutcracker (*a*) is the most dependent. It is also the most specialized: in addition to a sublingual pouch it has a long, sharp bill, with which it pries open pine cones to get at their seeds. Moreover, it is a strong flier capable of going a great distance burdened with seeds. The piñon jay (*b*) and Steller's jay (*c*) are intermediate in both behavior and anatomy. They carry seeds in their distensible esophagus, not in a specialized sublingual pouch. The scrub jay (*d*) is the least dependent on stored seeds and also the least specialized in its anatomy. It can carry seeds only in its mouth and its bill. In each drawing the seed-carrying capacity of the bird is indicated by a broken colored line.

little or no interocular transfer of information, the behavior is further evidence that they rely on their memory to recover their stores.

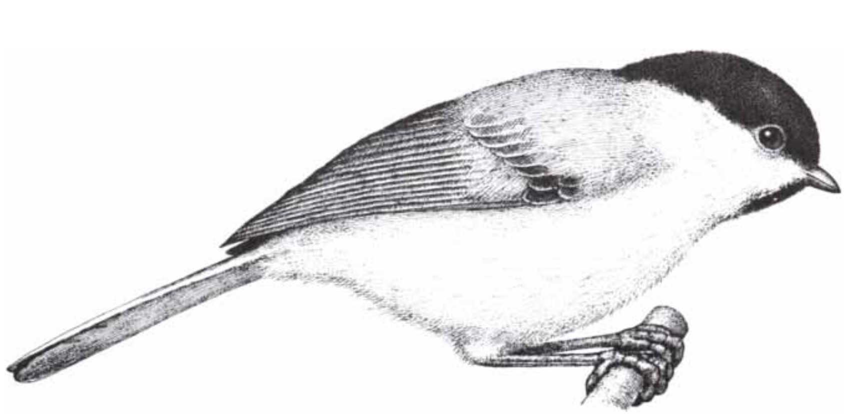
What about other food-storing birds? In the past few years several investigators have been studying birds such as Clark's nutcracker, which makes thousands of winter caches. Again evidence of memory has been sought both in the field and in the laboratory. Of course a field study attempting to discover such things as whether or not a Clark's nutcracker recovers its own caches is extremely difficult to do because of the large number of caches and the long interval between storage and recovery. Few people would want to wait several months near a nutcracker's cache, hoping to learn whether it was eventually dug up by the bird that buried it.

Diana Tomback of the University of Colorado at Denver reasoned, however, that much could be learned from the way Clark's nutcrackers search for caches. She took advantage of the fact that the birds leave a record of their searches in the form of beak marks in the snow and the earth where they dig. Moreover, successful searches can be identified by the presence of piñon-seed coats next to the holes from which seeds were taken. If nutcrackers search at random, successful probe holes and clusters of unsuccessful holes should be more or less evenly distributed over landscapes where the birds have searched. Tomback found, on the contrary, that unsuccessful probes were clumped around successful ones. The pattern indicated the nutcrackers were not searching for caches by trial and error. Furthermore, in the early spring (before rodents had found many of the caches) about two-thirds of the probes were successful, far more than one would expect of a random search.

Tomback's observations do not prove unequivocally that the birds could remember where their caches were. They might have smelled them (although again it is an unlikely possibility). Alternatively, they could have searched mainly in places of the kind that are likely to contain caches. Recent laboratory studies done by Stephen B. Vander Wall at Utah State University seem to rule out these possibilities. Vander Wall studied four captive Clark's nutcrackers in a large outdoor aviary. Two of the birds readily cached seeds in the sandy floor of the enclosure; the other two did not. Nevertheless, the nonhoarders did dig in the sand with their beaks and eat the buried seeds they happened to uncover. Vander Wall allowed the two hoarders to bury seeds in the aviary; then he allowed all four birds to search for seeds.

The results provided dramatic evidence of memory. When the nutcrackers that had hoarded seeds dug in the sand to find them, 70 percent of their probes were successful. When the nonhoarders dug, they found seeds in only 10 percent of their probes. To be sure, a success rate of 10 percent is substantially greater than one would expect for a bird probing completely at random. On the other hand, the two nutcrackers that cached seeds preferred to make their caches near prominent objects in the aviary, such as logs and rocks, and all four birds searched most often near those objects. This tendency accounts for the nonhoarders' success.

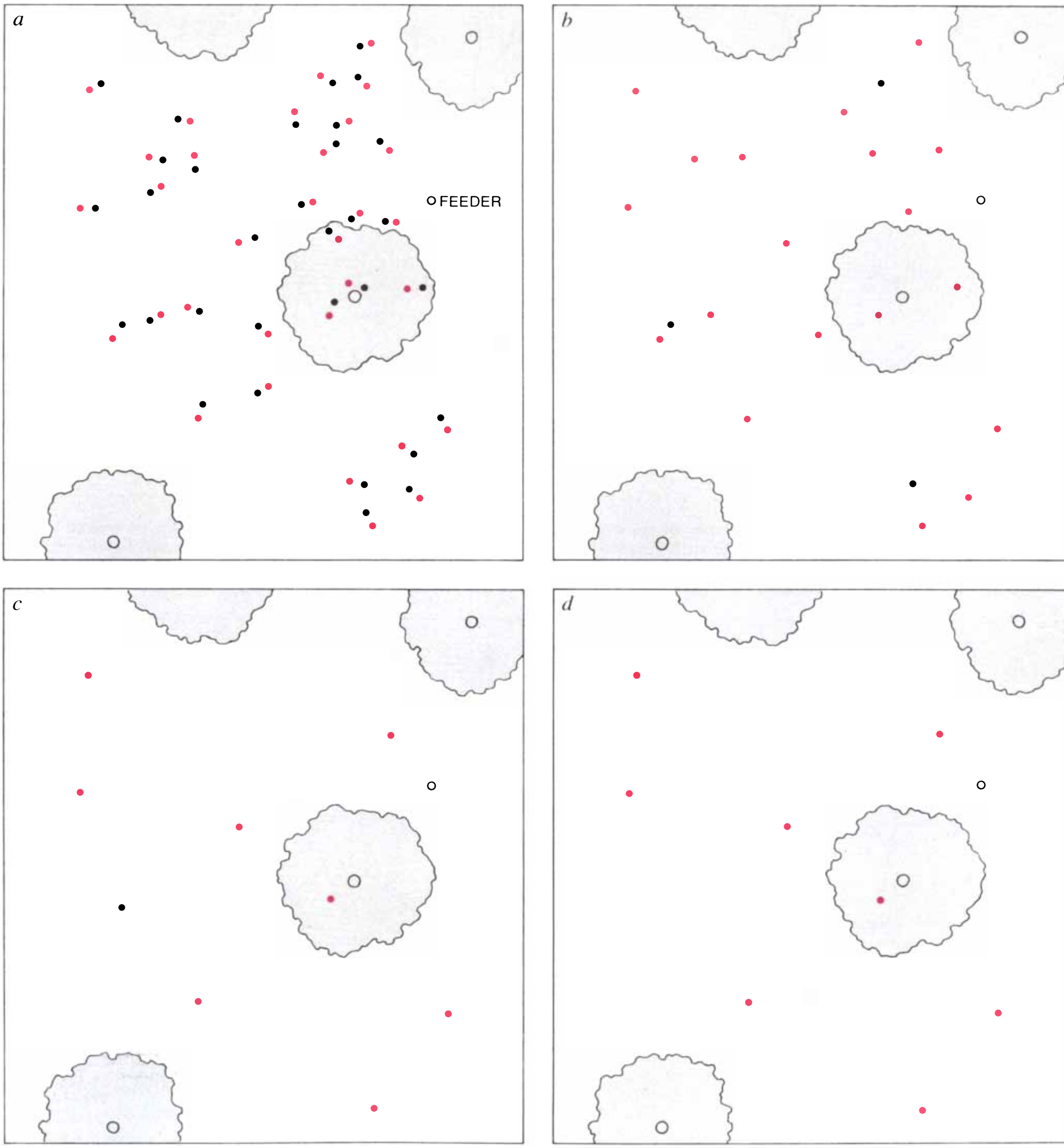
Perhaps nearby prominent objects



MARSH TIT is a small British bird (related to the North American chickadees) that caches seeds if there are more available than it needs immediately for food. Typically the marsh tit returns to draw on its caches within a few hours or at most a few days after it deposits them.

serve as cues for memory when the birds attempt to find their stores. To test this hypothesis Vander Wall covered the aviary floor with a plastic sheet, leaving exposed only an oval area where the sheet had been cut away. At each end of the oval he arranged four large objects such as rocks. The nutcrackers were allowed to make caches wherever they chose in the uncovered oval. Then one end of the oval was extended by 20 centimeters and the four objects at that end were moved in the direction of the lengthening so that they too were each 20 centimeters from their original position.

Would the nutcrackers be able to find caches at the altered end of the oval as well as the ones at the unaltered end, or would their probes be off by 20 centimeters? The latter proved to be the case. Toward the unaltered end of the oval the birds' success rate was high. Toward the altered end the success rate declined. Most of the probes at the altered end were within a few centimeters of the point that was "correct" with respect to the altered position of the nearest large object. The birds also made errors near the middle of the oval, but the errors

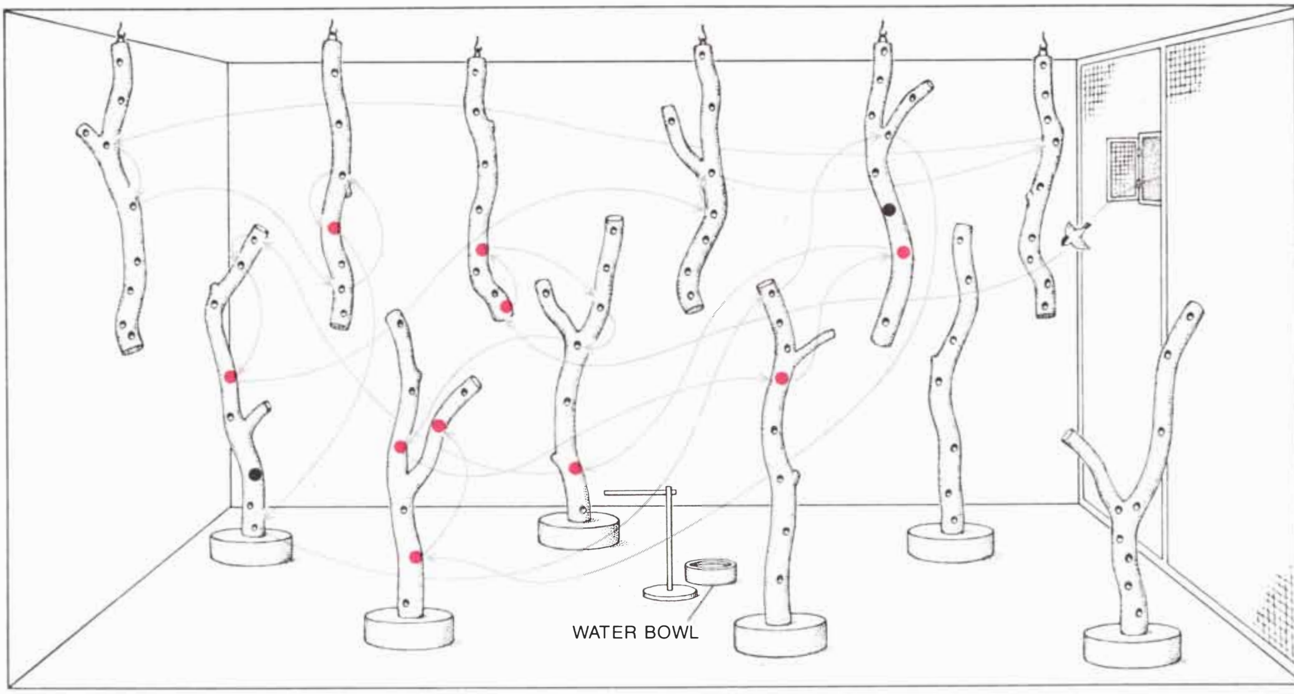


DISAPPEARANCE OF CACHES from Wytham Wood near Oxford was the first evidence confirming that marsh tits use memory to recover hoarded food. Richard J. Cowie, John R. Krebs and David F. Sherry of the University of Oxford coated 30 sunflower seeds with a radioactive substance, then employed a portable scintillation counter to find where a marsh tit had cached them (*black dots*) in Wytham Wood (*a*). One meter from each cache the investigators established a false cache (*colored dots*). Ten hours later (*b*) 19 false caches were undisturbed but only three of the bird's caches remained. The investigators made further checks after 19 hours (*c*) and after 30 hours (*d*).

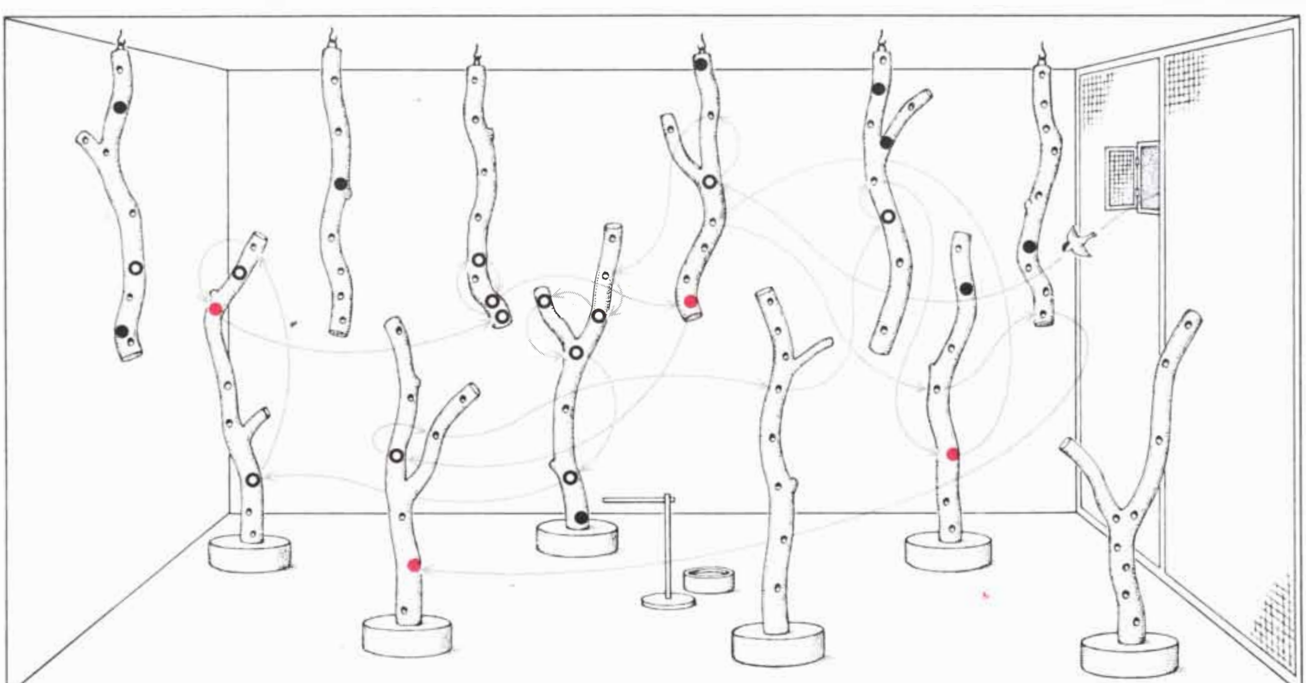
were less extreme. This finding suggests that the cache sites had been established with reference to both sets of objects.

For any species that disperses its stores and relies on memory to recover them one may ask whether the animals' memory enables them to benefit in other ways. That is, are food-storing species better in general at remembering where things are? An attempt to answer this question would require that animals be given tests of spatial memory that do not involve stored food. Such tests have not yet been done with food-storing birds, but thinking about how they might be done reveals some important points.

Among the laboratory techniques employed by psychologists who investigate memory in animals the closest analogue to the hoarding of food is a delayed-



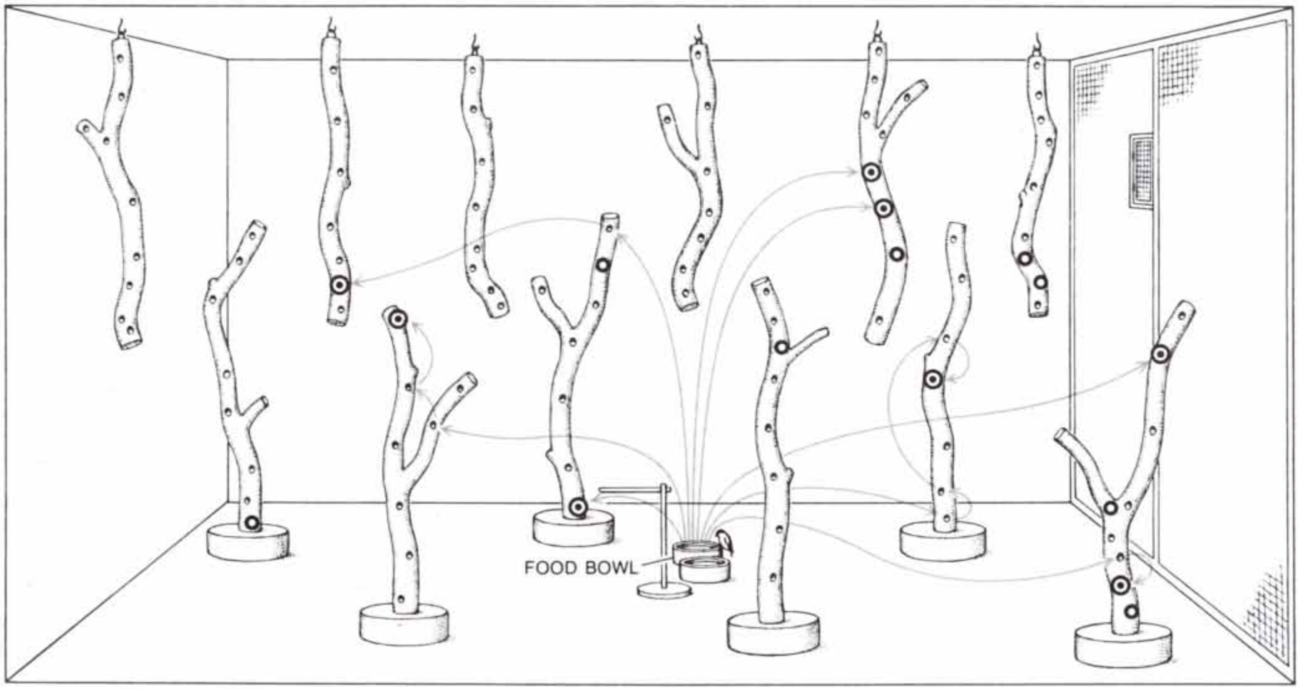
FOUR EXPERIMENTS testing the memory of marsh tits were done by the author in collaboration with Krebs. In each instance a bird stored seeds in some of the roughly 100 holes the experimenters had drilled in tree branches in the laboratory. Here a marsh tit hoarded 12 seeds (*dots*). Readmitted to the room two and a half hours later, it found 10 of the seeds (*colored dots*) by inspecting only 24 holes.



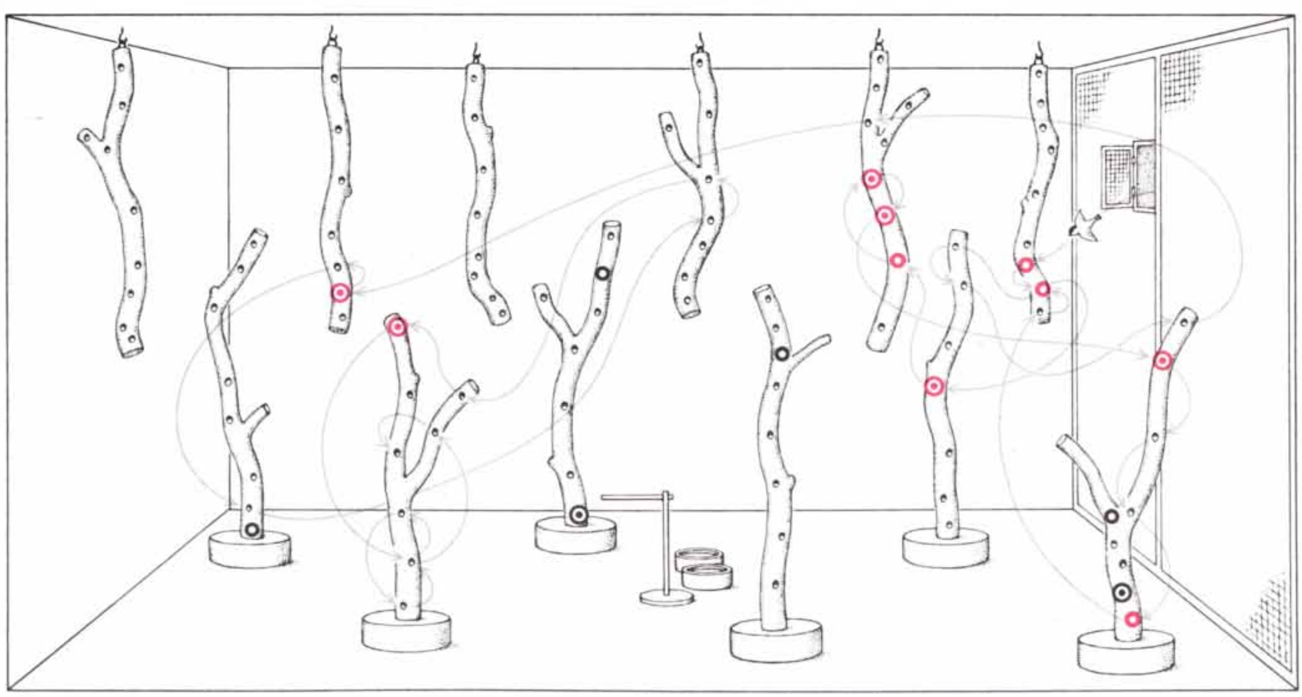
IN A SECOND EXPERIMENT a marsh tit stored 13 seeds (*open black circles*), but Krebs and the author then moved them to other holes (*dots*). Among the first 24 holes the bird inspected when it was returned to the room, 11 were holes it had employed. The experiment indicates that the bird was not detecting the seeds by a cue such as their odor. The bird did succeed in finding four seeds (*colored dots*).

response experiment. Here an animal is shown some visual cue, such as a flashing light or a geometric figure. Later it is offered a choice between the original cue and another one. It is rewarded for choosing the one it saw before. The performance of a laboratory rat or a pigeon typically falls to the level of chance if a few seconds or minutes pass between the first viewing and the second.

Recently Donald M. Wilkie and Russel Summers of the University of British Columbia trained pigeons in a variant of the delayed-response procedure that requires spatial memory. The work illustrates some of the differences between food hoarding and conventional laboratory studies of animal memory. In Wilkie and Summers' experiment a pigeon faced a three-by-three array of nine white disks that could be lighted in-



TWO BATCHES OF SEEDS were hoarded by a marsh tit in the third experiment. In choosing holes for the second batch (*black circles with dots*) it avoided the holes in which it had stored the first batch (*open black circles*). The experiment establishes that the bird was not merely visiting a particular set of holes whether it was storing seeds or recovering them. The marsh tit could remember where it had been.



RECOVERY OF BOTH BATCHES of seeds by the marsh tit that had hoarded them was the final experiment in the series. By inspecting 29 holes it found 10 seeds: four (*open colored circles*) from the first batch of eight and six (*colored circles with dots*) from the second batch of eight. Its first seven inspections of holes were all successes. In two instances the bird made the mistake of revisiting a hole.

dividually from behind. At the start of each trial one disk was lighted briefly. A few seconds later the disk was lighted again, along with another disk or several others. The pigeon was rewarded (it received a pellet of food) if it pecked the disk that had lighted up twice.

The pigeon's task seems to be quite similar to the task facing a food-storing bird. After all, a bird storing food is exposed to visual cues at the storage site, and in order to recover the food it must select the site in preference to similar sites nearby. Yet with only one cue to remember at a time, the pigeons in Wilkie and Summers' experiment appeared not to know which disk had been lighted only a few seconds after they saw it. In contrast, a Clark's nutcracker evidently remembers thousands of storage sites for months.

The pigeon's task did differ in a number of ways from the nutcracker's. The pigeon passively viewed the cue it would be called on to remember. The cues themselves were invariably within a few centimeters of one another and differed only in their position within the array. The food with which the pigeon was rewarded was not behind the cue itself. Any of these factors would be expected to make the pigeon's task more difficult than the task facing a food-hoarding bird. For instance, laboratory investigations done mainly with rats show that animals can remember for several hours the locations they actually visit.

Perhaps, therefore, the types of cues available at hoarding sites are what make the sites memorable. On that hypothesis a food-hoarding bird would not necessarily have a memory extraordinarily better than that of other animals. By adopting a life of hoarding food it would simply have put itself in circumstances where its memory is particularly serviceable. How could this be tested? Comparisons of memory between species that hoard food and species that do not are often compromised by differences in the motivations and the sensory and motor abilities of the species. Obviously the comparisons would be most significant if the species were closely related ones.

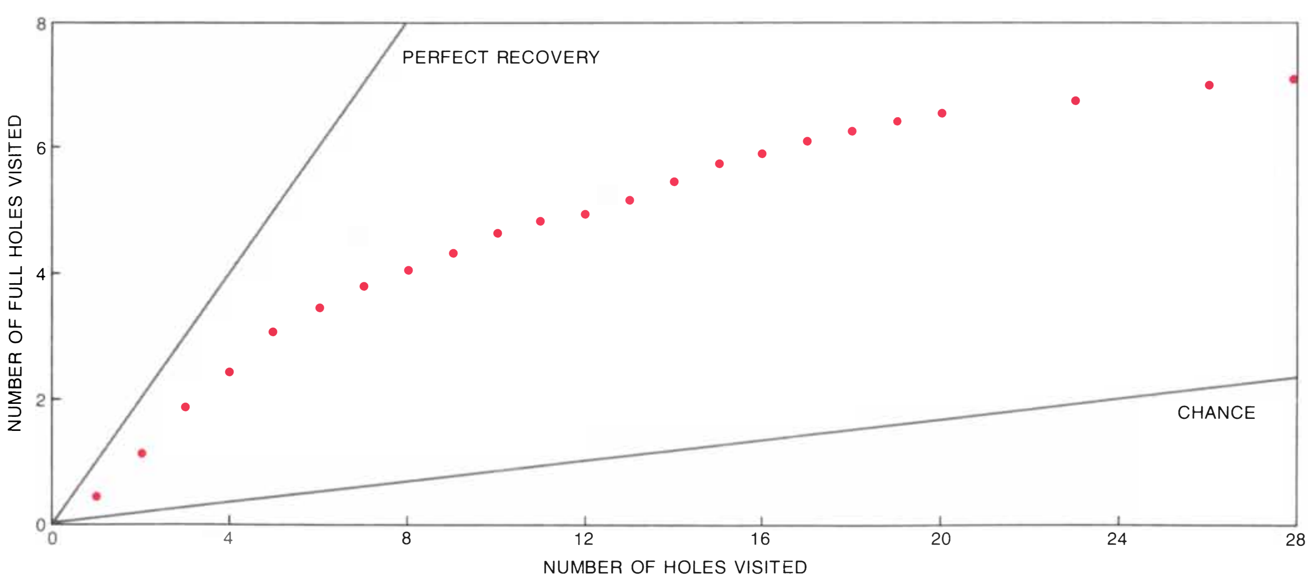
In this respect it is notable that among the corvids (the jays, crows and nutcrackers) and also among the tits some species store food whereas others do not. Among the tits, for example, crested tits and willow tits apparently store food in the fall and draw on it over the winter, much as Clark's nutcrackers do. Marsh tits store superabundant food for relatively short periods. Blue tits and great tits do not store food.

Among the corvids an example of several related species that differ in their tendency to hoard has been described by Vander Wall, working with Russell P. Balda of Northern Arizona University. Clark's nutcracker lives in the same areas of the western U.S. as three other members of the crow family do, namely the piñon jay, Steller's jay and the scrub jay. The four species differ greatly in their dependence on stored piñon seeds, and concomitantly they differ in how well they are able to harvest and transport the seeds.

Clark's nutcracker is the most dependent. By exploiting its stored seeds it is able to begin breeding as early as February, well before most other birds in the area. Clark's nutcracker is also the most specialized anatomically for harvesting and hoarding seeds. In addition to the sublingual pouch in which it carries large numbers of seeds, it has a long, sharp bill, with which it pries open green pine cones before the cones would have opened by themselves. Moreover, Clark's nutcracker is a strong flier. It can carry heavy loads great distances to suitable storage sites.

The scrub jay, in contrast, is the least specialized of the four species. It lacks the anatomical adaptations of a Clark's nutcracker, and it is much less reliant on stored seeds. The seeds it does store it caches entirely within the area where it lives: the area it defends as its own territory. The piñon jay and Steller's jay are intermediate in both anatomy and behavior. Is the spatial memory of the birds of these three species less highly developed than it is in a Clark's nutcracker? The answer awaits further work in the laboratory and the field.

The investigations of memory and food hoarding I have described are at the interface between zoological and psychological studies of animal behavior. Such research is increasing as more psychologists become convinced that it is productive to view laboratory studies of learning and memory as ways of analyzing how animals solve the problems confronting them in nature. At the same time zoologists are becoming interested in aspects of foraging behavior that clearly require a capacity for learning and memory. The joint efforts can therefore combine the zoologist's understanding of how behavior is adapted to the environment with the psychologist's conceptual tools and experimental techniques. The efforts are the beginning of exciting new developments in the study of memory.



MARSH TIT'S PERFORMANCE was charted by Krebs and the author. If the bird were searching at random for eight seeds in 100 holes, it would inspect about 12 holes to find each seed. If its memory were perfect, on the other hand, it would inspect just one hole to find each seed. The bird's performance was intermediate: in 12 trials it inspected an average of about 3.7 holes to find each seed.

Oscillating Chemical Reactions

Once thought to violate natural law, reactions where concentrations of key substances periodically rise and fall have now been designed deliberately. They may illuminate similar behavior in living systems

by Irving R. Epstein, Kenneth Kustin, Patrick De Kepper and Miklós Orbán

Oscillating or periodic phenomena are ubiquitous in physics, astronomy and biology. They range from the familiar motion of pendulums and the orbits of planets to the complex biological clocks that govern the daily and seasonal behavior of living organisms. Until fairly recently, however, chemists believed the reactions in their test tubes and beakers were singularly immune to the kind of periodic behavior so familiar in other fields of science. In fact, most chemists trained before 1950 would probably have said that for a mixture of simple inorganic substances to participate in a reaction that oscillated visibly and periodically would be a clear violation of an immutable law of nature. Even today it is usually assumed that chemical reactions are a one-way street: if two substances react to yield a third substance, it is expected that the reaction will continue steadily until the reactants are exhausted or until an equilibrium state is reached. It is not ordinarily expected that the concentrations of the intermediate products of the reaction will reach a certain level, then fall to a lower level, then rise and fall again repeatedly until at some point stable products resistant to further change result.

Although accounts of just such reactions appeared sporadically in the chemical literature of the late 19th century and the early 20th, they were dismissed by the large majority of chemists as being nonreproducible phenomena, attributable perhaps to extraneous processes such as corrosion or film formation occurring in the course of the reaction. In the past 25 years, however, the study of oscillating chemical reactions has finally won respectability and now constitutes one of the most rapidly growing branches of chemistry. The study promises to yield fresh insights into chemical dynamics and the mechanism of catalysis, and possibly into baffling periodic phenomena observed in biology and geology. Whereas the original small group of oscillating chemical reactions was found by accident, it is now possible to describe conditions conducive to oscillation and to mount a systematic search for new oscillating systems. It is a search that has already been rewarded.

The reluctance of chemists to accept the reality of oscillating reactions can be traced chiefly to the second law of thermodynamics. In its best-known formulation, achieved by the 19th-century German physicist Rudolf Clausius, the second law states that the entropy, or randomness, of the universe tends to increase. Applied to chemical reactions the principle requires that a chemical system, in the absence of external inputs of matter or energy, must continuously approach an ultimate equilibrium state. That is, if *A* goes to *B,* it must do so throughout the course of the reaction without detours back to *A* along the way. Reactions where this rule appears to be violated must result, it was thought, either from poorly controlled experimental conditions or even from deliberate deception, since they would constitute a kind of chemical perpetual-motion machine.

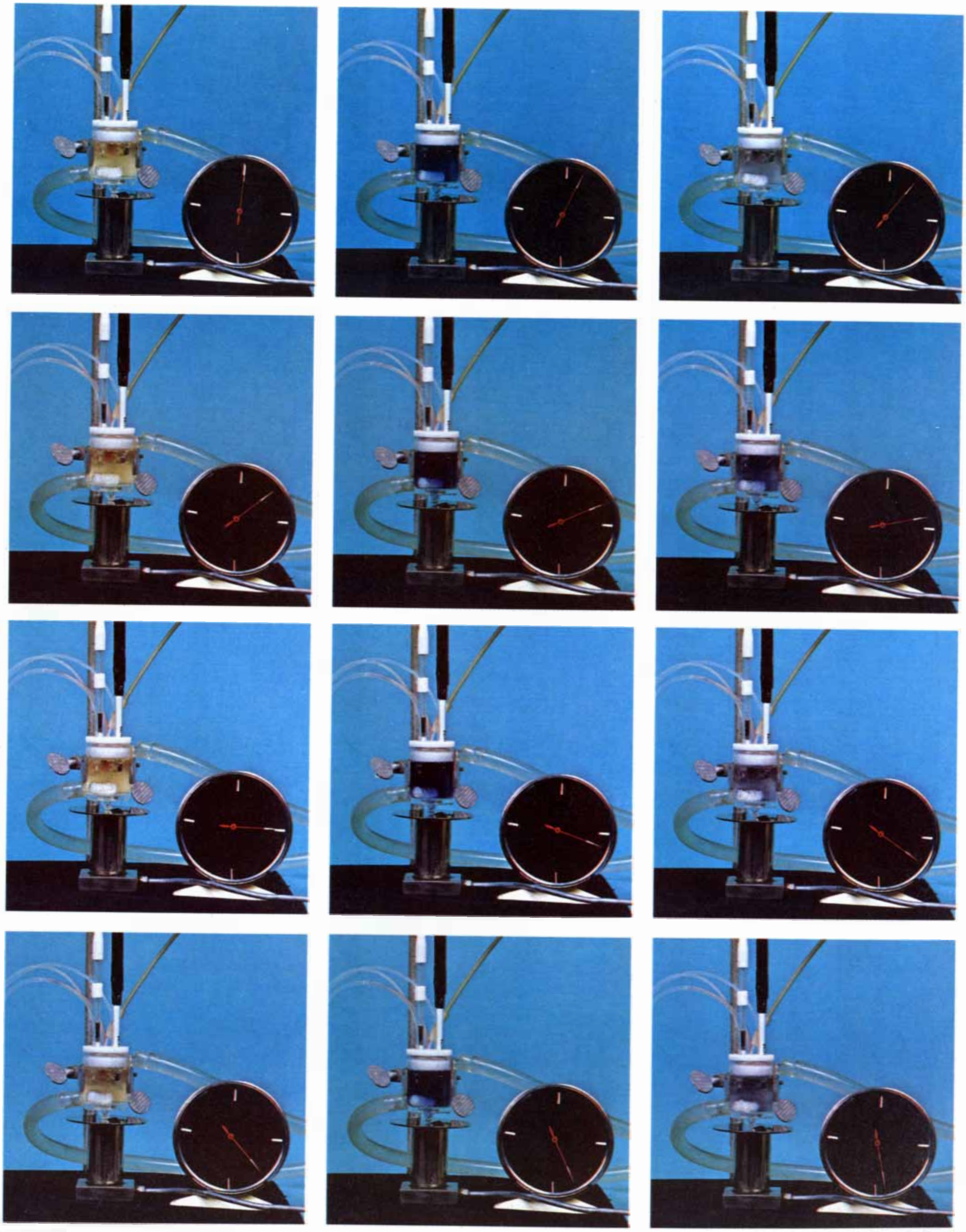
One can therefore imagine the indifference that greeted the report of an oscillating reaction published in 1921 by William C. Bray of the University of California at Berkeley. In studying the role of iodate, the ion of oxidized iodine (IO_3^-), in catalyzing the decomposition of hydrogen peroxide to water and oxygen, Bray noticed that under certain conditions the rate of oxygen production and the concentration of iodine in the solution changed periodically. Bray's observation received scant attention for the next half century. The few scholarly papers that did appear on the Bray reaction were devoted largely to explaining away the oscillations as artifacts due to dust particles or impurities rather than to attempting to examine the phenomenon further.

As we shall see, iodine, the iodide ion (I^-) and the iodate ion play a major role in several of the newly designed oscillating systems. The modern era in the study of oscillating chemical reactions can be dated from an accidental discovery in 1958 by the Russian chemist B. P. Belousov. He noticed that if citric acid and sulfuric acid are dissolved in water with potassium bromate and a cerium salt, the color of the mixture changes periodically from colorless to pale yellow. By 1958 some physical chemists, at least, were prepared to take the report seriously, both because the Belousov reaction was easy to repeat and in part because the 19th-century concepts of thermodynamics had been extended considerably in the years following World War II.

A leader in this effort was Ilya Prigogine of the Free University of Brussels, who recognized that the classical thermodynamics of Clausius required not only that systems be isolated from their environment but also that they be near their equilibrium state. For systems far from equilibrium, either because a reaction is only in its early stages or because the system is "open," hence subject to a flux of energy or matter from the outside, Prigogine and his co-workers developed the concept of an irreversible thermodynamics. For this work Prigogine received the 1977 Nobel prize in chemistry.

In systems that are far from equilibrium a number of new phenomena, described as dissipative structures, can arise. Such structures include periodic oscillations in the concentrations of intermediate species in a chemical reaction; the initial reactants and final products, however, are not subject to oscillation. Living systems are the most interesting and varied examples of open, far-from-equilibrium oscillators. They are maintained in a nonequilibrium state by the input of reactants (nutrients) from the external environment and the output of products (wastes). If either of these fluxes is stopped, the organism and its oscillations must die.

In 1958 the implications of Prigogine's irreversible thermodynamics were understood by only a handful of chemists. Moreover, Belousov's discovery went almost unnoticed because it was



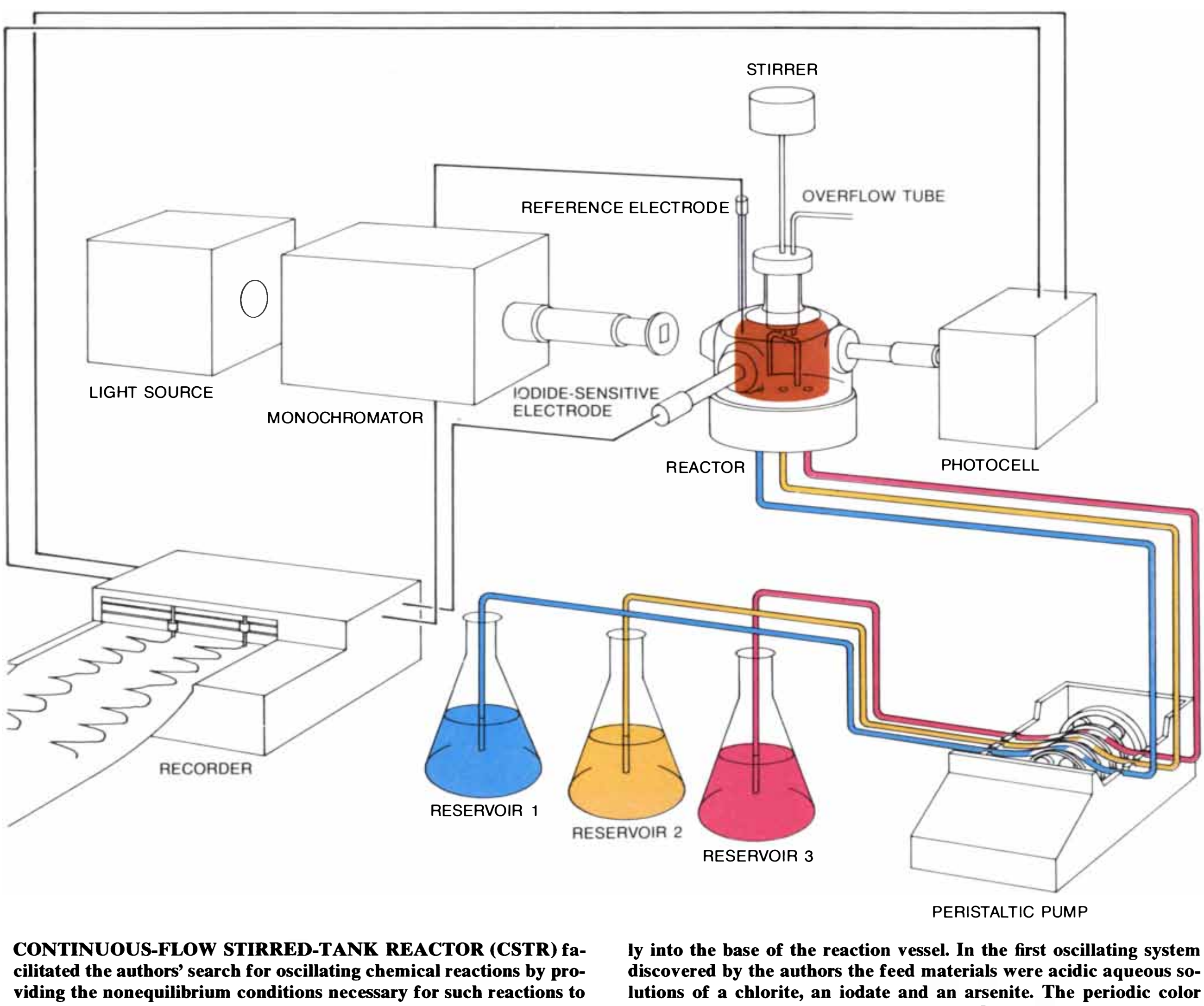
OSCILLATING CHEMICAL SYSTEM is depicted in the authors' laboratory at Brandeis University. The reaction is proceeding in a glass vessel into which three solutions are being pumped at a steady rate: potassium iodate, a solution of perchloric acid and hydrogen peroxide, and a solution of malonic acid containing manganous sulfate monohydrate. Starch serves as an indicator, forming a dark blue complex in the presence of iodine. The interval between successive photographs, beginning at the top left and reading from left to right, is just under three seconds (as is indicated by the second hand on the clock). This particular reaction system was devised by two California high school teachers, Thomas S. Briggs and Warren C. Rauscher. The reaction has been modified to run under continuous-flow conditions.

published in an obscure Russian collection of technical abstracts in radiation medicine. Belousov's oscillating reaction had important advantages over the earlier Bray system. It ran at room temperature and it did not generate noxious products. Furthermore, the oscillations were clearly visible as the cerium changed back and forth from an oxidized (ceric) state, which is yellow, to a less oxidized (cerous) one, which is colorless.

A few years after its publication Belousov's reaction caught the attention of A. M. Zhabotinsky of the Institute of Biological Physics near Moscow. He modified Belousov's recipe somewhat (for example by replacing the cerium reagent with an iron reagent that gave a more distinctive red-blue color change) and began a systematic study of the reaction, which is now commonly called the BZ (for Belousov-Zhabotinsky) reaction. In the early 1960's Zhabotinsky published a large volume of experimental results, including the finding that if a thin layer of initially uniform red BZ solution is allowed to remain undisturbed, blue dots soon appear and grow into a striking array of ring-shaped structures.

Knowledge of the BZ reaction (and of Prigogine's irreversible thermodynamics) spread rapidly in the late 1960's and early 1970's. Although chemists all over the world were fascinated by the BZ phenomenon, the great majority still regarded it as a curiosity, more useful for astonishing students in lectures than as a subject for serious study. In a short time many variants of the BZ reaction were developed by replacing one or more of the components with a closely related chemical species to yield a slower, faster or differently colored oscillator. The most visually impressive of the variants was a hybrid of the BZ and Bray systems devised in 1973 by two San Francisco high school teachers, Thomas S. Briggs and Warren C. Rauscher. Their system, which involves hydrogen peroxide, potassium iodate, perchloric acid, malonic acid, manganese sulfate and starch, changes from colorless to gold to blue and back again. (The detailed recipe is given in the *Scientific American* department "The Amateur Scientist" for July, 1978.)

Although the BZ reaction and its descendants had become well known by the late 1960's through lecture demonstrations, the nature of the oscillatory phenomenon remained a mystery. A chemist understands a reaction only when he can write a mechanism for it. A mechanism is a set of component reactions, called elementary steps, each of which describes an actual encounter between molecules. The familiar "stoichiometric" equation that chemists usually write to describe a reaction shows only the net result, not how the reaction actually proceeds on a molecular scale. For example, when acetylene (C_2H_2)



CONTINUOUS-FLOW STIRRED-TANK REACTOR (CSTR) facilitated the authors' search for oscillating chemical reactions by providing the nonequilibrium conditions necessary for such reactions to take place. The illustration shows a typical experimental arrangement. The reactants, held in three reservoirs, are pumped continuously into the base of the reaction vessel. In the first oscillating system discovered by the authors the feed materials were acidic aqueous solutions of a chlorite, an iodate and an arsenite. The periodic color changes are monitored by the photocell. Oscillations in the concentration of particular ions (here iodide) are sensed by an electrode.

combines with oxygen (O_2) in a welder's torch to form carbon dioxide (CO_2) and water (H_2O), the overall reaction can be written $2C_2H_2 + 5O_2 \rightarrow 4CO_2 + 2H_2O$.

It is wildly improbable, of course, that two acetylene molecules and five oxygen molecules would ever collide simultaneously and fly apart to give four molecules of carbon dioxide and two of water. The reaction actually takes place through a sequence of elementary steps, each one involving a collision between two molecules or the breakdown of a single molecule, often creating or destroying intermediate species that do not appear at all in the equation of the net reaction. Unraveling such a mechanism is an enormous challenge calling for considerable ingenuity, the analysis of much experimental data and in many instances extensive computer simulation.

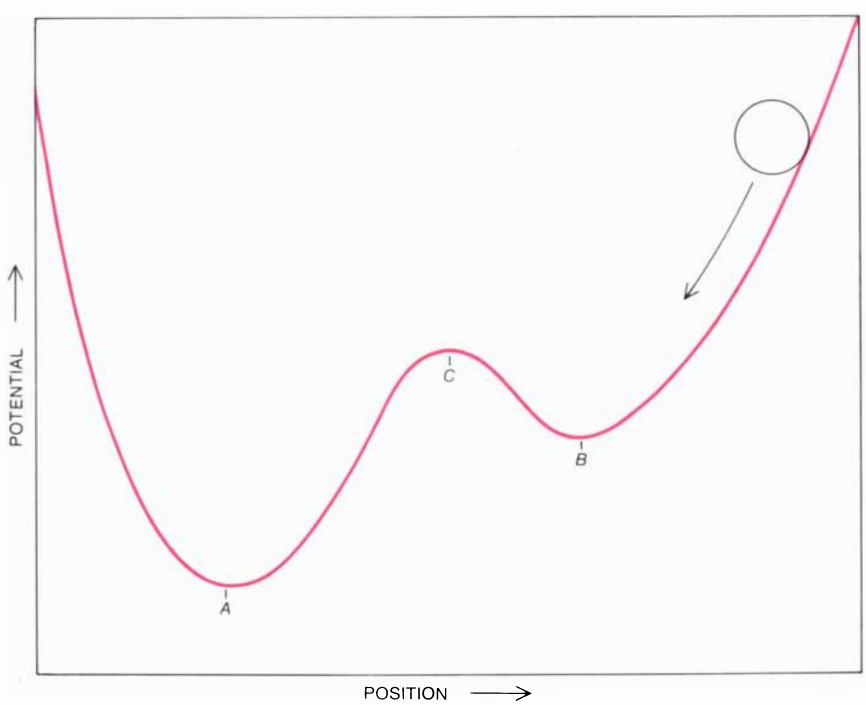
The formidable task of constructing a mechanism for the oscillating BZ reaction was undertaken in the early 1970's by Richard M. Noyes of the University of Oregon in collaboration with Richard J. Field (who is now at the University of Montana) and Endre Körös, a visiting scholar from Eötvös University in Budapest. By 1972 they had conceived a scheme consisting of 18 elementary steps, involving some 20 different chemical species, that appeared capable of explaining the oscillations. Two years later a detailed computer simulation by Noyes, Field and David Edelson of Bell Laboratories confirmed that the mechanism does predic: oscillation. Further studies have shown that the mechanism accounts for the development of spatial structures as well. Noyes has likened their approach in unraveling the mechanism to the "method of [Sherlock] Holmes: 'When all other contingencies fail, whatever remains, *however improbable,* must be the truth.'"

With the solution of the BZ mechanism, oscillating chemical reactions became a respectable and exciting area for those interested in reaction mechanisms and chemical dynamics. Various hypotheses were advanced on the conditions required for oscillation in a chemical reaction, and there was optimism in some circles that an understanding of chemical oscillation might throw new light on periodic behavior in living organisms. No one could succeed, however, in defining a set of necessary and sufficient conditions for chemical oscillation. As the decade of the 1970's drew to a close the number of fundamentally different chemical oscillators known was discouragingly small; all had been discovered by accident; only one was well understood, and no one knew how to find any more.

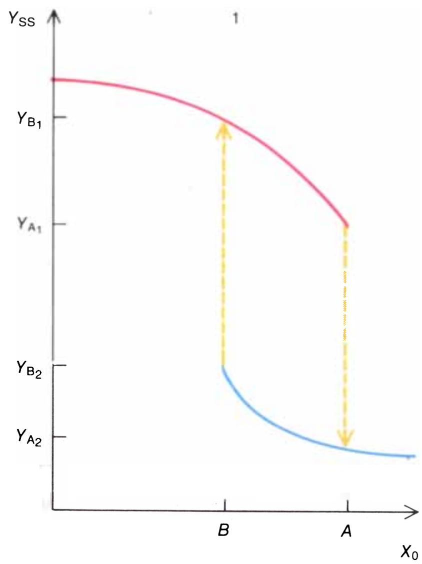
Although biological systems were known to be a fertile source of preparations that exhibit oscillation, they tend

1	$2H^+ + Br^- + BrO_3^- \rightleftharpoons HOBr + HBrO_2$
2	$H^+ + HBrO_2 + Br^- \rightleftharpoons 2HOBr$
3	$HOBr + Br^- + H^+ \rightleftharpoons Br_2 + H_2O$
4	$CH_2(COOH)_2 \rightleftharpoons (OH)_2C{=}CHCOOH$
5	$Br_2 + (OH)_2C{=}CHCOOH \rightleftharpoons H^+ + Br^- + BrCH(COOH)_2$
6	$HBrO_2 + BrO_3^- + H^+ \rightleftharpoons 2BrO_2 + H_2O$
7	$BrO_2 + Ce^{3+} + H^+ \rightleftharpoons Ce^{4+} + HBrO_2$
8	$Ce^{4+} + BrO_2 + H_2O \rightleftharpoons BrO_3^- + 2H^+ + Ce^{3+}$
9	$2HBrO_2 \rightleftharpoons HOBr + BrO_3^- + H^+$
10	$Ce^{4+} + CH_2(COOH)_2 \rightleftharpoons CH(COOH)_2 + Ce^{3+} + H^+$
11	$CH(COOH)_2 + BrCH(COOH)_2 + H_2O \rightleftharpoons Br^- + CH_2(COOH)_2 + HOC(COOH)_2 + H^+$
12	$Ce^{4+} + BrCH(COOH)_2 + H_2O \rightleftharpoons Br^- + HOC(COOH)_2 + Ce^{3+} + 2H^+$
13	$2HOC(COOH)_2 \rightleftharpoons HOCH(COOH)_2 + O{=}CHCOOH + CO_2$
14	$Ce^{4+} + HOCH(COOH)_2 \rightleftharpoons HOC(COOH)_2 + Ce^{3+} + H^+$
15	$Ce^{4+} + O{=}CHCOOH \rightleftharpoons O{=}CCOOH + Ce^{3+} + H^+$
16	$2O{=}CCOOH + H_2O \rightleftharpoons O{=}CHCOOH + HCOOH + CO_2$
17	$Br_2 + HCOOH \rightarrow 2Br^- + CO_2 + 2H^+$
18	$2CH(COOH)_2 + H_2O \rightleftharpoons CH_2(COOH)_2 + HOCH(COOH)_2$

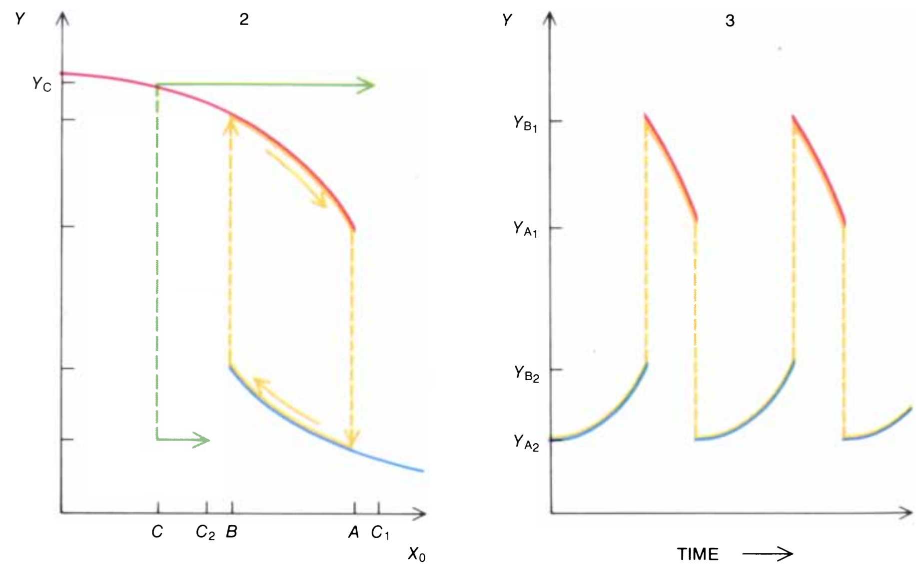
COMPUTER SIMULATION OF OSCILLATION in the Belousov-Zhabotinsky reaction involves 18 elementary steps and 21 different chemical species. The reaction is named for two Russian chemists: B. P. Belousov, who discovered it, and A. M. Zhabotinsky, who improved on it. The starting materials are three inorganic substances, bromate ions (BrO_3^-), bromide ions (Br^-) and cerous ions (Ce^{3+}), and one organic substance, malonic acid ($CH_2(COOH)_2$). A sulfuric acid medium supplies hydrogen ions (H^+). The products of the reaction are carbon dioxide (CO_2), formic acid (HCOOH) and bromomalonic acid ($BrCH(COOH)_2$). As the cerium oscillates between the ceric (Ce^{4+}) and cerous (Ce^{3+}) states the solution is alternately yellow and clear. The reactions shown in color involve only inorganic species and are better understood than the reactions shown in black, which involve species derived from malonic acid. This mechanism was worked out at the University of Oregon by Richard M. Noyes together with his postdoctoral associate Richard J. Field and Endre Körös of Eötvös University in Budapest.



BISTABLE CHEMICAL SYSTEM can be represented by a mechanical analogy. A ball rolling in a potential well with two low points, *A* and *B*, may come to rest in either one. *A* and *B* thus represent stable steady states of the system; *C* is an unstable steady state. Although a ball may come to rest at *C*, the smallest perturbation will topple it into either *A* or *B*. Here the variable is the position of the ball; the external constraint is the shape of the potential well.



HYSTERESIS LOOP is depicted in a bistable chemical system operating in a continuous-flow stirred-tank reactor. As the concentration X_0 of substance X flowing into the reactor is increased from a low value the steady-state value Y_{SS} of species Y in the reactor decreases smoothly along the red curve until at the critical point A it drops suddenly from Y_{A_1} to Y_{A_2}. If X_0 is further increased, Y_{SS} again falls but along the blue curve. If X_0 is decreased, Y_{SS} retraces the blue curve past point A until a second critical point B is reached. Here Y_{SS} jumps from Y_{B_2} to Y_{B_1}. Between A and B either steady state can exist.



OSCILLATION CAN RESULT if a suitable substance Z is added to a bistable chemical system that exhibits hysteresis. If substance Z reacts with substance Y at a low rate to yield substance X, the effective concentration of substance X_0 is modified so that the system "sees" a higher value of X than is actually fed from the reservoir (*left*). The effect of a given amount of Z on X_0 (*green arrows*) will be greater on the red branch of the hysteresis loop (where there is more Y) than on the blue branch. Imagine that the reaction is started with a flow of X equal to C. One now adds enough Z, an amount Z_0, so that the effective X_0 is equal to C_1 when the system is on the red branch and equal to C_2 when it is on the blue. In the absence of Z the flow $X_0 = C$ would yield a steady-state value Y_c. With the addition of Z_0, however, the system follows the red curve as the effective value of X is slowly raised because of the reaction of Y with Z. The system tries to reach the value of C_1 but on reaching the value A it drops abruptly from the red branch and undergoes a rapid transition to the blue branch. There the system "sees" a value of $X_0 = C_2$, and so it drifts to the left along the blue curve in an attempt to match the new effective value of X_0. Again before it reaches that goal, at $X_0 = B$ it suddenly jumps to the red

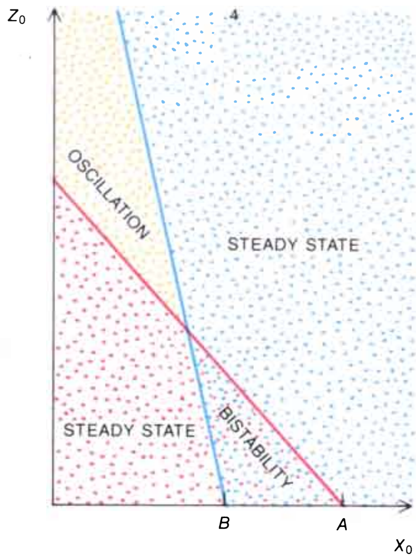
to defy the kind of mechanistic analysis that leads to generalization. Such oscillating systems can be prepared, for example, by extracting the chemical contents of yeast cells and separating them from the cell walls and other extraneous structures. The resulting "soup," when it is fed with appropriate nutrients, exhibits oscillation in the concentrations of hydrogen ions and of the molecule nicotinamide adenine dinucleotide, which in its reduced form (NADH) transports electrons in the cell's normal energy cycle. At the Johnson Research Foundation, Britton Chance and his colleagues have shown that many such biological oscillators involve glycolysis, the process by which cells generate energy by metabolizing sugar in the absence of oxygen. Since such reactions are catalyzed by an array of enzymes, they are daunting in their complexity.

In 1979 two of us (Epstein and Kustin) decided the time had come to try to devise a systematic procedure for designing a chemical oscillator. Where, however, should one start? We began by identifying three conditions that were known to be either necessary for chemical oscillation or conducive to it. The first condition is that chemical systems can oscillate only if they are far from equilibrium. The second is feedback: some product of a step in the reaction sequence must exert an influence on its own rate of formation. The third is that the chemical system must exhibit bistability. Under the same set of external conditions, referred to as constraints, the system must be able to exist in two different stable steady states.

The requirement that the system be maintained far from equilibrium could be met by having the reaction proceed in an apparatus well known to chemical engineers: the continuous-flow stirred-tank reactor (CSTR). In the mid-1970's Adolphe Pacault and his co-workers at the Paul Pascal Research Center in Bordeaux had adapted the CSTR to the study of oscillating chemical reactions. When one of us (De Kepper) left Bordeaux to join our group at Brandeis University in 1980, it was a natural step to build such a reactor. The arrival from Hungary of another of us (Orbán) who had had extensive experience with bromate oscillators completed the group.

A common type of feedback encountered in living systems is autocatalysis, in which the rate at which a substance is produced increases with its concentration. The concept extends to populations of organisms: the number of organisms added to a population in a given interval is almost always proportional to the number of individuals already in the population. (This observation underlies Thomas Malthus' prediction of 1798 that populations tend to outrun food supplies.) In chemistry autocatalytic systems are rare but by no means extraordinary. As early as 1920 Alfred J. Lotka of Johns Hopkins University showed that a simple scheme of two coupled autocatalytic reactions would give rise to oscillation. Although the Lotka mechanism has not proved applicable to any actual chemical reaction, it has guided the thinking of many chemists and has been of considerable utility in describing the oscillations of predator and prey populations in ecological systems.

Let us now consider the role of bistability, or the existence of two different stable steady states, in an oscillating chemical system. By a steady state is meant a condition of the system in which all the variables, such as the concentration of each chemical species, have attained constant values. The steady state is stable if it can adjust to a small change in a variable, say the addition of a drop of acid, and recover without being transformed into a new state. If the slightest change causes a transition to a different state of the system, the original state is said to be unstable. For example, a small ball resting in the bottom of a bowl is in a stable state. A ball



branch and the cycle is completed (*yellow arrows*). X_0 now has the effective value of C_1 again, so that the system moves once more to the right. Repetition of this sequence yields a periodic oscillation in the value of Y (*middle*). The cross-shaped phase diagram (*right*) shows how raising the value of Z_0 can shift the system from bistability to oscillation. If Z equals 0, the system exhibits two steady-state conditions: the "red" state for a small X_0 and the "blue" state for a large X_0. When X_0 lies between A and B, the system is bistable. As Z_0 is increased the region of bistability narrows until at a critical value oscillation begins.

balanced on the rim of the bowl is in an unstable state. In a continuous-flow stirred-tank reactor the external constraints that control the steady-state values in the reaction system are usually the temperature of the thermostatic bath, the concentrations of the chemicals in the reservoirs feeding the reactor and their rates of flow.

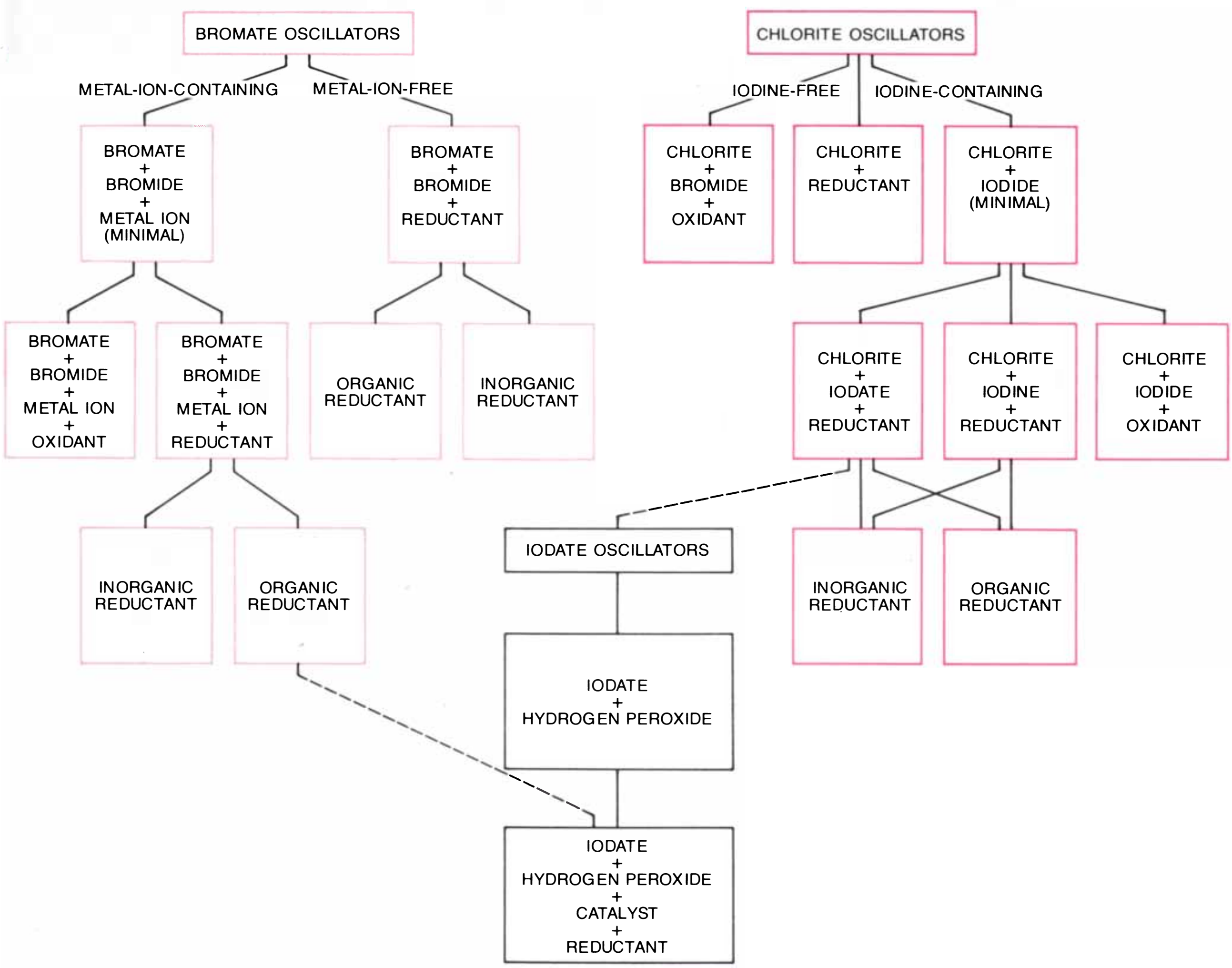
If the external constraints on a bistable chemical system are changed, a most peculiar phenomenon known as hysteresis may ensue. The phenomenon is familiar in magnetism, where it takes the form of the "hysteresis loop." When a piece of iron is subjected to increasing magnetizing force, the iron eventually attains full magnetic saturation. If the magnetizing force is then reduced to zero, the iron retains some of its magnetization. In order to reduce the magnetization to zero the iron sample must be subjected to a magnetizing force in the opposite direction. The complete cycle of magnetization and demagnetization takes the form of an *S*-shaped loop, broad in the middle and pointed at the ends.

An analogous phenomenon can occur in a bistable chemical system. In 1979 a calculation by Jacques Boissonade and one of us (De Kepper) at the Paul Pascal Research Center suggested that oscillations might be induced by adding another chemical to a system capable of hysteresis. The behavior of such a system can be represented in a cross-shaped phase diagram, which shows that if the added substance exerts significantly different effects on the two branches of the system's stable states, oscillations are likely to develop.

Our plan of attack for designing a chemical oscillator included the following four steps: find an autocatalytic system, run the reaction in a continuous-flow reactor, vary the conditions until a region of bistability is found and then introduce another substance capable of affecting the two branches of bistability differently and thereby of inducing oscillations. Our search of the chemical literature for autocatalytic reactions uncovered several promising ones. Two of them seemed particularly attractive because they shared iodine as a common intermediate. One reaction involved iodate ions and arsenite; a second involved iodide ions and chlorite. Conceivably one reaction might have the desired perturbing influence on the other. The iodate-arsenite reaction soon proved to be bistable. When chlorite was introduced, the composite system began to oscillate almost immediately. The first systematic effort to design a new chemical oscillator was a success.

By varying the components that enter the continuous-flow reactor we have since expanded the initial chlorite-iodate-arsenite oscillator into a family of two dozen related oscillating systems, all having the chlorite ion in common. We can outline a taxonomy of chlorite oscillators, and we are beginning to see how they can be related to oscillators of the bromate and iodate families.

Although bromate oscillators were among the first to be discovered and are far better understood mechanistically than chlorite oscillators, the range of chlorite systems we have found by deliberate design is much broader. Two factors, one chemical and the other historical, account for this seeming paradox. Chemically each family of oscillators has a minimal, or simplest, member from which the others can be derived by adding other substances. The minimal chlorite oscillator, chlorite plus iodide, oscillates over a rather wide range of conditions. In contrast the minimal bromate oscillator, bromate plus bromide plus a metal ion, functions only under the most precisely defined constraints. In fact, even after the existence of the



MAJOR TYPES OF CHEMICAL OSCILLATORS are grouped according to a tentative taxonomy into three main families. The original Belousov-Zhabotinsky oscillator and its many variants are found in the branch of the bromate family that has metal ions. The bromate oscillator in the branch that is free of metal ions is the work of Körös and one of the authors (Orbán) at Eötvös University. All the chlorite oscillators have been discovered by the authors at Brandeis. All the bromate systems involving inorganic reductants were developed at Brandeis under continuous-flow conditions. The iodate family has the fewest members. The first system listed under that heading is the oscillating reaction reported in 1921 by William C. Bray of the University of California at Berkeley and largely ignored for 50 years. The second system in the iodate family is a hybrid of the Belousov-Zhabotinsky and Bray systems. Broken lines connect systems that have certain reactants in common. A reductant is a substance that donates electrons to another substance; an oxidant removes electrons.

minimal bromate oscillator had been predicted by one investigator, Kedma Bar-Eli of Tel-Aviv University, the search for it was fruitless until two of us (Orbán and Epstein) succeeded in finding it by following our systematic approach based on the cross-shaped phase diagram.

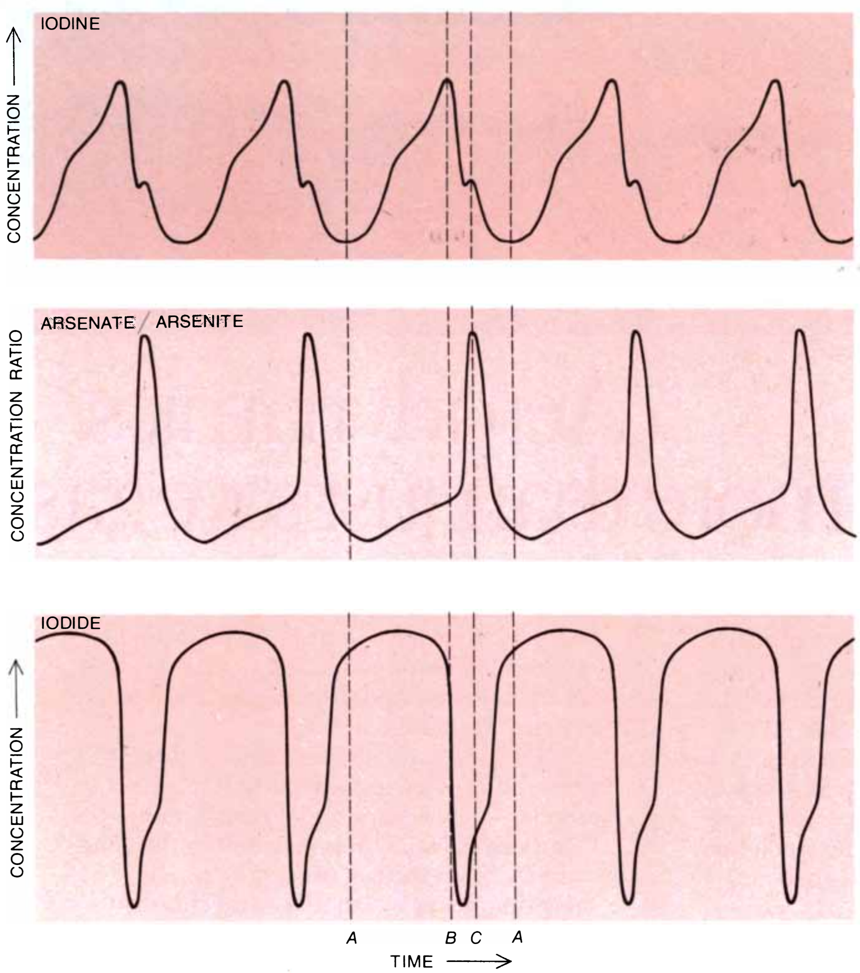
From the historical point of view the first bromate oscillators were discovered before continuous-flow reactors came into service. The bromate systems oscillated readily under batch (zero flow) conditions. Chlorite oscillators, like the vast majority of new systems, show periodic behavior only under continuous-flow conditions because the flow is needed to maintain them far from equilibrium. A system that oscillates under batch conditions will nearly always oscillate under continuous flow, but the converse is rarely true. Most bromate oscillators were originally sought under batch conditions; chlorite oscillators were originally sought only under flow conditions. Accordingly the variety of chlorite systems discovered soon exceeded the variety of bromate systems. (We subsequently found, however, several chlorite oscillators that work under batch conditions.) One might say that the very ease of oscillation of the first bromate systems hindered the search for more of them. It is only quite recently that continuous-flow reactors have been intensively applied to the search for new bromate oscillators and to the study of the third (and so far the smallest) family of oscillators: the iodate systems.

One of the aims of designing a new chemical oscillator was to have a system whose mechanism could be deciphered and compared with that of previously examined oscillating reactions. The dissection of the family of oscillating chlorite systems is now well under way. We have completed studies of the rates of several component processes and thus have several pieces of the puzzle; a complete description of the oscillating scheme now appears to be a matter of turning groups of pieces around until they lock together in just the right way.

We can already give a schematic view of how our first oscillating system functions, the one whose inputs are solutions of either potassium or sodium salts in the form of a chlorite, an iodate and an arsenite. Only four principal overall reactions are involved, each consisting of several elementary steps. Two of them involve the reaction of chlorite with either iodide or iodine, yielding chloride and iodine in the first case and chloride and iodate in the second. In the third reaction iodate and iodide react to form iodine. In the fourth iodine and arsenite form iodide and arsenate. Variations in the rates of these four reactions give rise to periodic oscillations in the concentrations of iodine and iodate and in the ratio of arsenate to arsenite [*see illustration at left*].

Chemical oscillation, although fascinating in itself, is related to several other phenomena of at least equal interest. One is the formation of spatial structures in an initially homogeneous medium. Such structures develop, for example, in one of the oscillators we have discovered where the ingredients are chlorite, iodide and malonic acid. When starch is present as an indicator, the solution is initially a uniform purple, because of a complex formed by iodide, iodine and starch. As the reaction proceeds white dots appear and grow into rings and sets of concentric rings, which annihilate one another when they collide. One observer compared the emergence of white dots against the purple background to seeing stars come out. The appearance of such geometric forms in a previously undifferentiated mass is highly suggestive of the kind of process that may enable embryonic cells in animal organisms to sort themselves out into individual types destined to become blood, brain or bone.

Another related phenomenon, whose



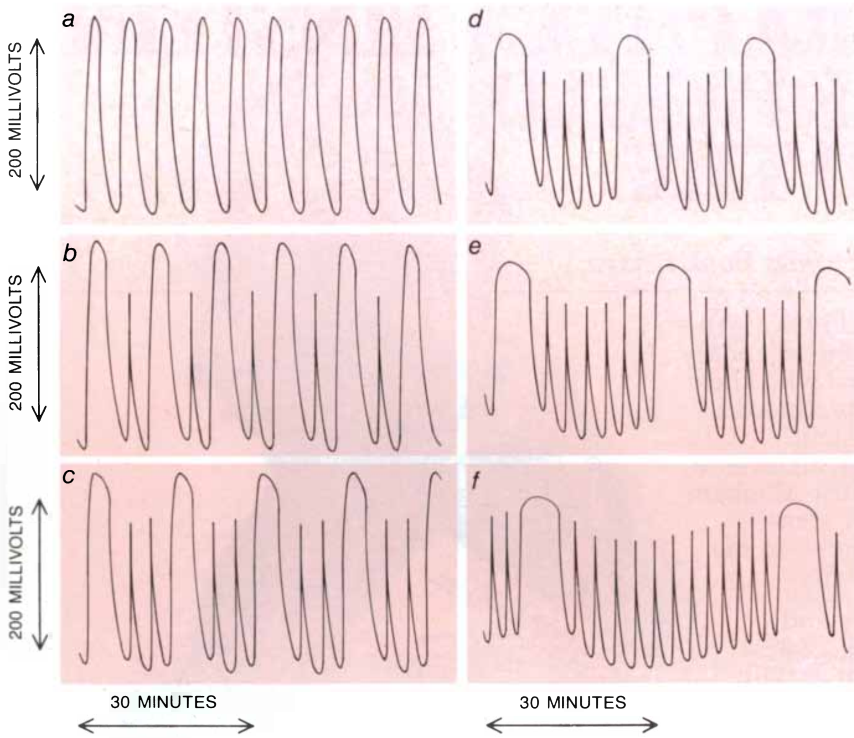
1 CHLORITE + IODIDE → CHLORIDE + IODINE 3 IODATE + IODIDE → IODINE
2 CHLORITE + IODINE → CHLORIDE + IODATE 4 IODINE + ARSENITE → IODIDE + ARSENATE

PATTERN OF OSCILLATIONS in the first oscillating system devised by the authors can be understood qualitatively from the interplay of four main reactions. The oscillating system is supplied with a uniform and balanced flow of three salt solutions: a chlorite, an arsenite and an iodate. Within the continuous-flow reactor the concentrations of iodine and iodate and the ratio of arsenate to arsenite oscillate as is shown by the three curves. At point *A* iodide and arsenite are high whereas iodine and arsenate are low, so that the predominant reactions are *1* and *3*. These reactions sharply increase the concentration of iodine and begin to decrease the concentration of iodide, bringing the system to point *B* and shutting down reactions *1* and *3*. Reactions *2* and *4* now come into play, depleting the iodine and converting arsenite into arsenate. The system reaches point *C*, where arsenite and iodide are at or near their minimum levels. Arsenite is replenished by the input to the reactor and arsenate is depleted by the output. Reaction *4* regenerates iodide while consuming the iodine, thereby returning the system to *A*.

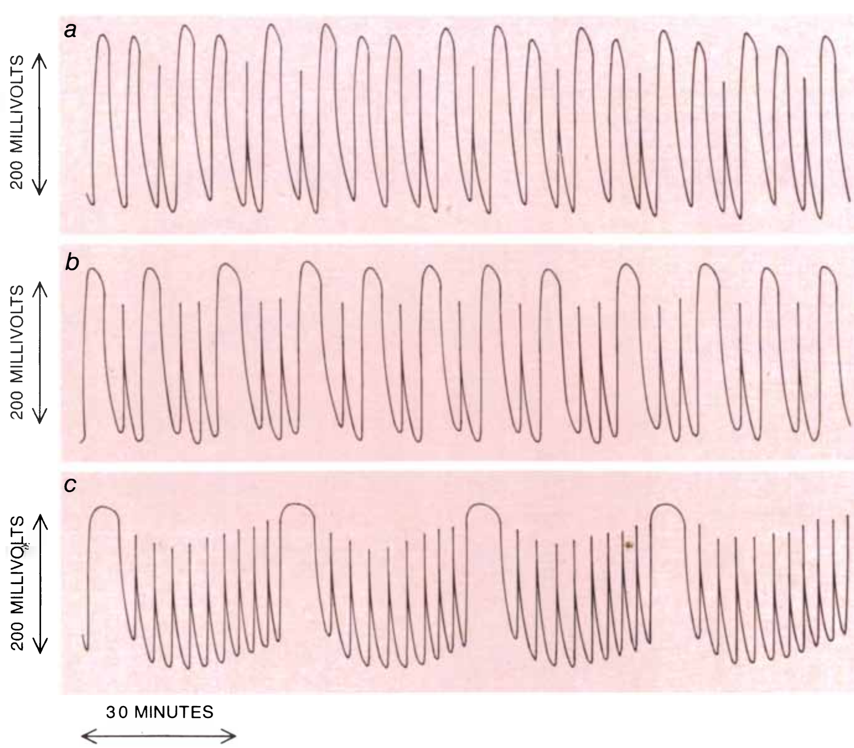
existence has aroused considerable controversy in the chemical community, much as the reality of oscillating reactions was questioned a quarter of a century ago, is "chemical chaos." The term refers to the behavior of a reaction in which concentrations neither attain constant values nor oscillate periodically but rather increase and decrease in an apparently random and unpredictable way. We have seen such fluctuations in the BZ system and in one of our new chlorite oscillators, when both are run in a continuous-flow reactor with careful monitoring. They are of particular interest to mathematicians because if they were genuinely inherent in the dynamics of the reaction, they would be physical examples of the mathematical objects known as strange attractors. Briefly and somewhat crudely, strange attractors can turn up when certain equations are iterated, or solved repeatedly, with the result that one iteration serves as the input for the next. In a system of equations with a strange attractor the path mapped out by the successive solutions appears to vary unpredictably from one cycle to the next. (Douglas R. Hofstadter explored the topic in his *Scientific American* department "Metamagical Themas" for November, 1981.)

Skeptical chemists have suggested that chemical chaos may represent nothing more than uncontrolled experimental fluctuations, for example in temperature or flow rate, which randomly push a system from one mode of complex periodic oscillation to another. Although the jury is still out, recent careful experimental and theoretical work by Harry L. Swinney and Jack Turner of the University of Texas at Austin, by Jean Claude Roux, Christian Vidal and their group at the Paul Pascal Research Center and by J. L. Hudson of the University of Virginia all strongly suggest that chemical chaos is a genuine phenomenon inherent in the dynamics of at least some chemical oscillators. The implications of such intrinsic chaos are not yet understood. Might it be, as Henry Adams once observed, that "chaos often breeds life, when order breeds habit"?

Now that chemical oscillators have shown themselves to be not only consistent with the laws of nature but also amenable to systematic design and description, they are coming to play an increasingly prominent role in several areas of science. The switching from one state to another that characterizes an oscillating reaction may hold the key to understanding regulatory processes in living cells, such as the mechanisms that turn on and off the copying of a strand of DNA or the contraction of a muscle. The same forces that create colored rings and layers in unstirred chemical oscillators may also be responsible for the spacing of the rings of Saturn and



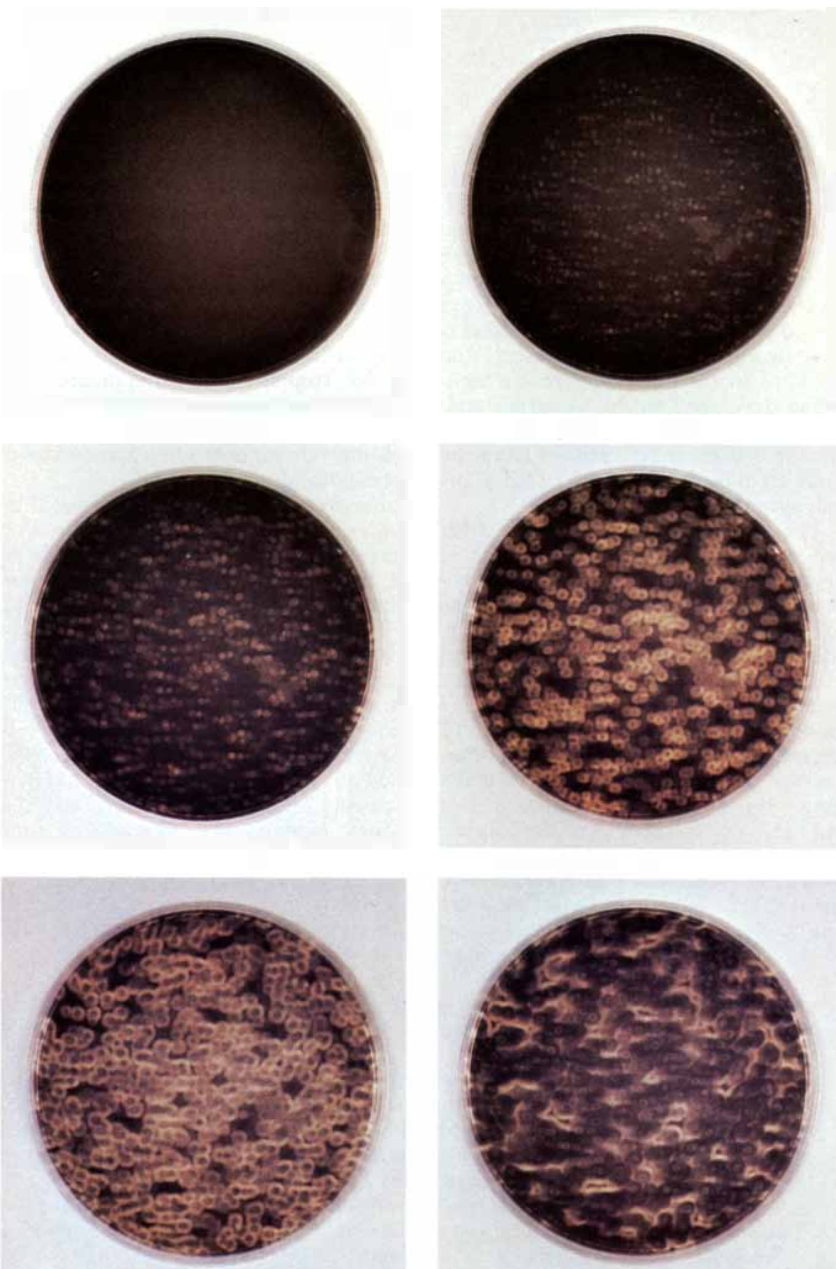
PERIODIC OSCILLATIONS in the chlorite-thiosulfate reaction exhibit a variety of patterns when the input rate to a continuous-flow reactor is altered systematically. Each cycle has one large oscillation and *n* small ones, where *n* can vary from 0 to 16 and is the same for all cycles.



APERIODIC, OR CHAOTIC, OSCILLATION is sometimes observed in a narrow range of flow rates between rates that yield complex periodic behavior. One form of chaos consists of an apparently random mixture of two types of periodic oscillation, one with *n* and one with $n + 1$ small oscillations for each large oscillation. One possible explanation is that uncontrolled variations in the flow rate randomly flip the system back and forth from a rate that gives *n*-type oscillation to one that gives ($n + 1$)-type. The alternative, which is favored by the authors, is that chaotic oscillation is intrinsic in the dynamics of the chemical reaction at certain flow rates.

the periodic striations found in certain rock formations, which are not readily explained by conventional geologic processes.

It also seems plausible that many and perhaps all of the catalytic reactions fundamental to the chemical industry proceed in an oscillatory manner. The catalyst may alternate between two forms on a time scale too short to be detected by standard analytic procedures. The study of oscillating reactions can provide a testing ground for theories of catalysis and, more generally, for theories of the dynamics of chemical reactions. Any model that can successfully predict the complex behavior of oscillating systems should be able to deal readily with the simpler phenomena observed in less exotic reactions. As more chemical oscillators are found or are deliberately designed, one may hope to obtain key insights into a variety of phenomena, ultimately including those clocks and carousels that operate within us all.



SPATIAL STRUCTURE develops in an oscillator based on chlorite, iodide and malonic acid discovered by the authors. The solution, placed in a shallow glass dish, is initially the uniform purple (*top left*) of a complex of iodine, iodide and a starch indicator. As the reaction proceeds white dots appear, then grow into rings and sets of concentric rings, which annihilate one another on colliding. The last picture was made about 90 seconds after the start of the reaction.

Autopsy

Postmortem examination contributes in many ways to medical knowledge and the planning of health care. Yet in recent years the rate of autopsy in the U.S. has declined to only 15 percent

by Stephen A. Geller

The only way to be certain what caused a person's death is to do an autopsy, or postmortem examination, of the body. The autopsy often reveals more than the cause of death. It can uncover diseases that went unrecognized in life, and it is a means of checking on the efficacy of medicine, surgery or other therapy that the person received in life for diseases that had been recognized. Yet barely 15 percent of the deaths in the U.S. are followed by autopsy now, whereas in the years after World War II the autopsy rate was nearly 50 percent. What accounts for the decline? Is the present low rate of autopsy detrimental to the cause of good medical care?

The word autopsy is derived from the Greek for seeing for oneself. As it is usually done, the procedure entails three kinds of seeing: a careful examination of the exterior of the body, a dissection and examination of the major organs and a study under the microscope of tissues from the organs. In the external examination, which includes examination of the orifices of the body, the pathologist is checking for evidence of abnormalities of any kind. They include growths, recent or old wounds and recent or old surgical scars.

Dissection begins with a large incision in the shape of a *Y*. The upper branches of the *Y* start at the front of each shoulder, cross below the nipples and join at the bottom of the sternum (the breastbone). The incision is continued down the middle of the abdomen to the pubis. In the abdomen it extends through the layers of the abdominal wall to enter the peritoneal cavity, which contains many of the major organs including the liver, the kidneys, the stomach, the intestines and the genitourinary system. The tissues are folded to the side to expose the abdominal cavity. Similarly, the sternum is removed to expose the cavities of the chest, which contains the heart, the lungs and the thymus. Each organ and system of organs is examined in place for evidence of abnormalities. As a rule the intestinal tract, from the duodenum to the rectum, is removed before the other organs are dissected.

The dissection of organs is usually done by one of two methods developed in the 19th century by the German physicians Rudolf Virchow and Friedrich Albert von Zenker. In Virchow's method the organs are removed from the body one by one and each organ is dissected and examined separately. Von Zenker's method, derived from a technique developed by his Austrian contemporary Karl von Rokitansky, calls for the organs to be removed from the body en masse and then dissected as organ systems.

The von Zenker approach is best for the examination and preservation of connections, normal or abnormal, between organs and organ systems. One example of the approach is that all the organs, beginning with the trachea and extending to the anorectum, the urethra and (in females) the vagina are removed en bloc. Once they are out the organs of the chest are separated from the abdominal and pelvic organs. The heart and the major blood vessels associated with it are examined in relation to the respiratory system and then are dissected free. When the abdominal organs are studied, the liver, the pancreas, the gall bladder and the bile duct are examined in terms of their relation to one another and to the stomach, the small intestine, the spleen and the associated blood vessels before they are separated and dissected individually.

The brain can be examined by cutting into the back of the scalp and part of the skull. It is studied in position, freed from its attachments to blood vessels and nerves and removed (by severing the lower part of the medulla oblongata). The spinal cord is removed by freeing it from its lateral attachments and extracting it through the skull or by cutting through the vertebral bodies.

In these steps as in other parts of the autopsy the method can be modified to meet the needs of the clinical problem to be solved. The breasts and the muscles can be studied without additional incisions of the skin. The face and the limbs are usually not dissected except with special permission in cases where the information obtained may be of particular clinical significance. All the incisions are of the surgical type and can be covered with clothing for a funeral.

After the organs have been studied as systems they are examined individually. Weights and measurements are recorded for comparison with standard values. Incisions are made in each organ to identify fine details. Grossly observable features are recorded in a standardized terminology.

For examination by microscope thin pieces of tissue from various organs are embedded in wax. Extremely thin slices of tissue are cut from these paraffin blocks, placed on glass slides, stained with various dyes and examined with the light microscope. For special problems an electron microscope may be brought into play. Additional techniques may include attempts to recover microorganisms that may have contributed to disease and chemical analyses designed to detect a wide range of substances perhaps made internally during life or taken in from outside.

The entire procedure takes from two to four hours, depending on whether the autopsy is total or limited, meaning that it deals with only one organ or a few organs that are thought to be of particular relevance to the case. Probably about 75 percent of the autopsies done in the U.S. are total. When the autopsy is finished, the organs are returned to the body (except for those that by permission are being donated to living people or are serving as a source of pharmacologic substances such as human growth hormone) and the incisions are sewn up. The final step is the compilation of a report of the findings.

The autopsy has been an integral part of medical practice for more than 100 years, but its purposes have changed somewhat. At one time the autopsy's main role was to identify the basis of disease and the cause of death, and often that is still the chief reason for doing one. When someone dies unattended or without medical care, the autopsy is the only means of understanding the terminal events.

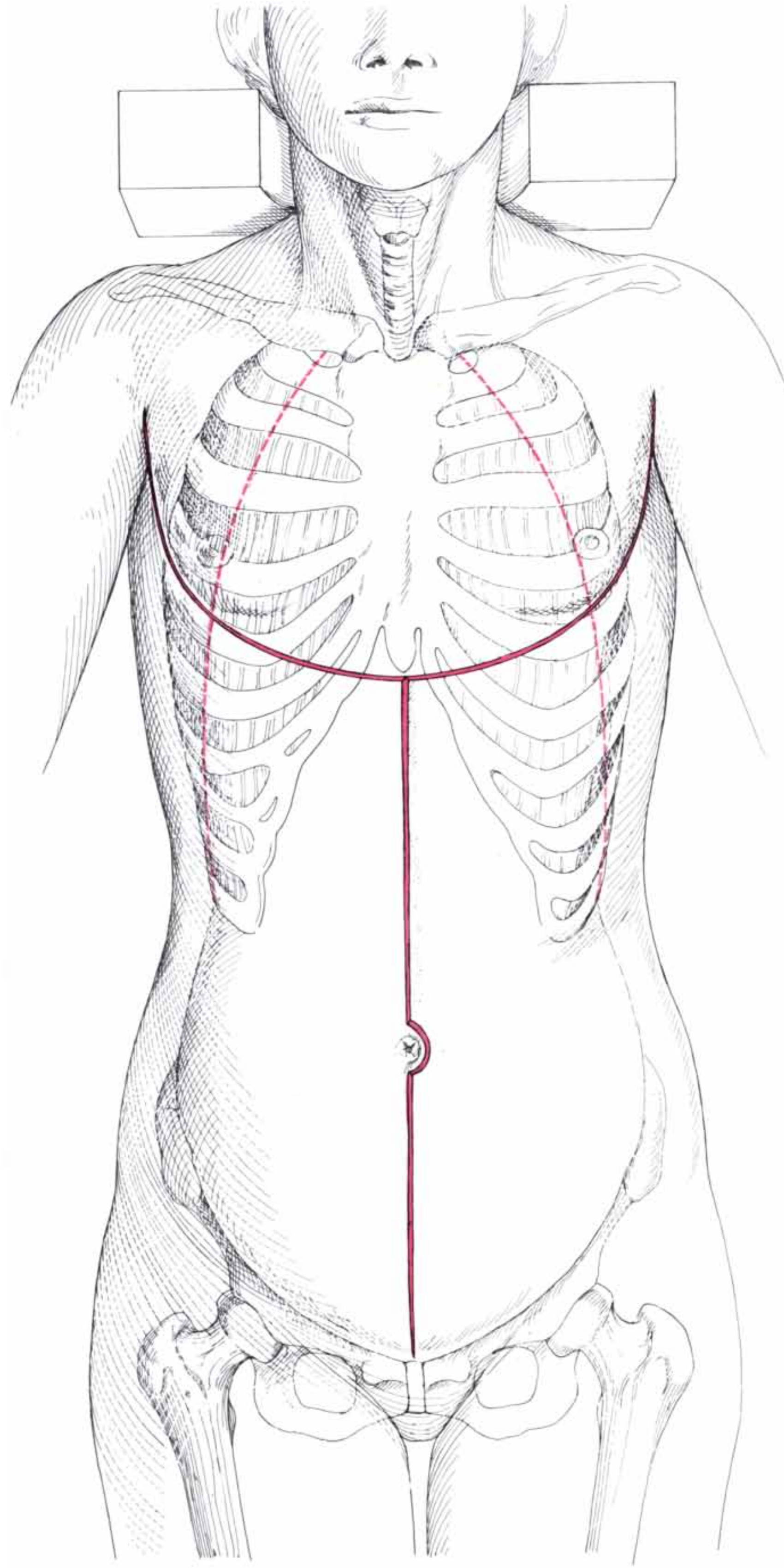
In most of the larger cities of the U.S. a medical examiner has the responsibility for reviewing the deaths that take place outside the hospital. Almost without exception, however, the office of the medical examiner is overburdened with deaths related to possible violations of law, and people whose death seems to have been due to natural causes are not autopsied.

For many of the deaths that occur in a hospital the principal disease has been diagnosed before death. In these cases the autopsy serves the particularly important function of a quality-control device. It monitors the accuracy of interpretation of diagnostic tests, determines the efficacy and toxicity of therapeutic agents, allows for the detection of conditions that may have been important but were either clinically inapparent or obscured by the most prominent disease and helps to monitor the influence of environmental factors on physiological processes. In addition the autopsy is a highly effective method for the education of medical students and the continuing education of physicians on the variations in human disease.

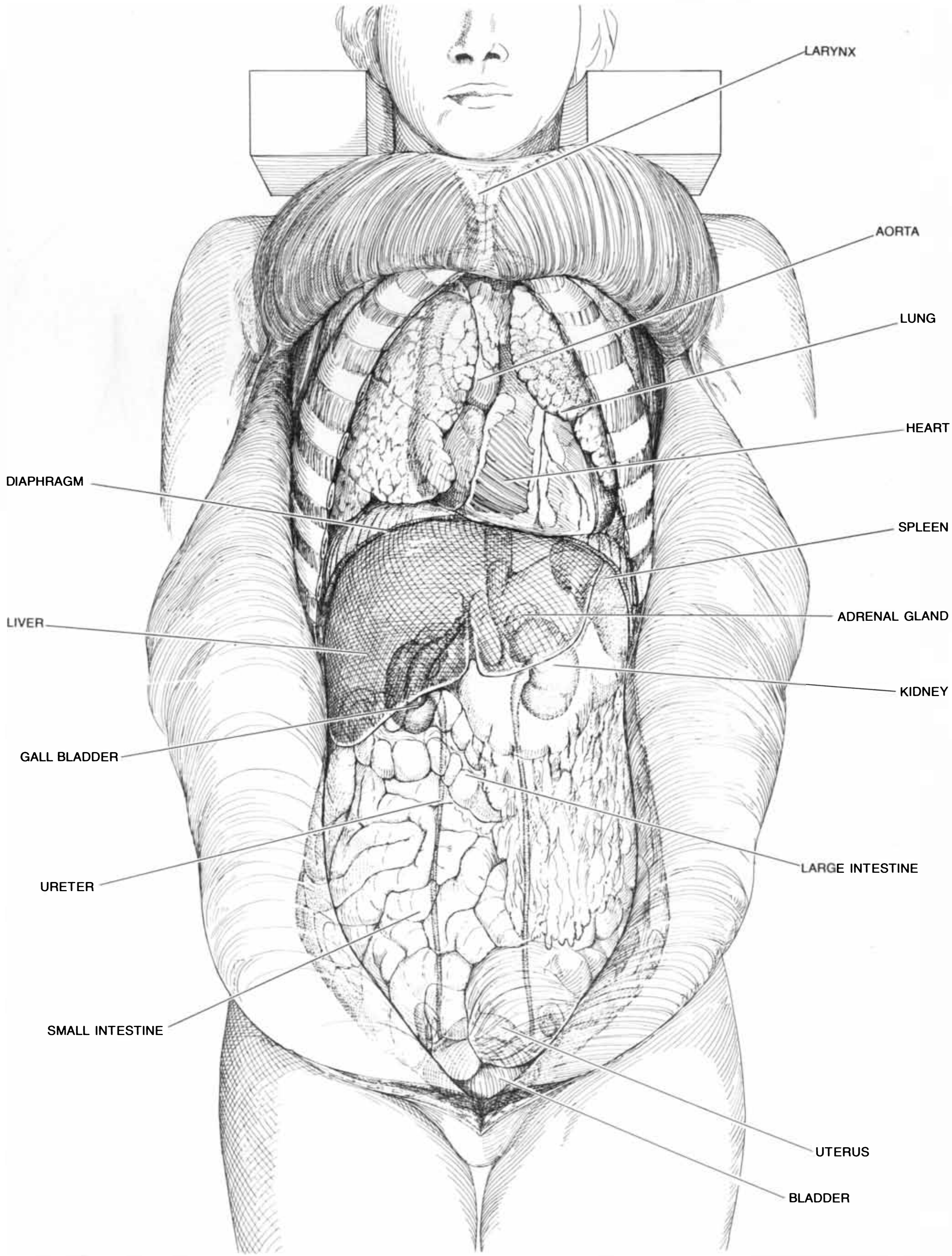
A measure of the value of autopsies is that even with the many recent advances in medicine the major diagnosis turns out to have been wrong in as many as 40 percent of the people autopsied. Even in the best medical centers the major disease is shown to have been incorrectly diagnosed in more than 15 percent of the autopsied cases and the immediate cause of death to have been clinically unrecognized in more than 40 percent.

Since most patients who die are not examined after death, there are no reliable figures on the overall accuracy of medical diagnosis. One indication that the rate of error is consistently high even in major medical centers was provided by the late Edward A. Gall, who was professor of pathology and chairman of the department of pathology of the University of Cincinnati School of Medicine. He documented the remarkable similarity of the percentage of errors in diagnosis, as disclosed by autopsy, at two major institutions over the 50-year period from 1912 to 1962.

A clear corollary of this situation is that the health statistics dealing with the causes of death are not useful. Indeed, the inaccuracy of death certificates has been well documented. Linda W. Engel and her co-workers at the National Cancer Institute assessed the accuracy of certification of the underlying cause of death in 2,557 cases autopsied in vari-



FIRST INCISION in the standard autopsy procedure is made following an external examination of the body for abnormalities. The incision (*color*) is in the form of a *Y* and is made to enable the pathologist to examine the organs of the chest, the abdomen and the pelvis. Then (*broken line*) the ribs are cut so that the breastbone can be removed to expose the chest cavities.

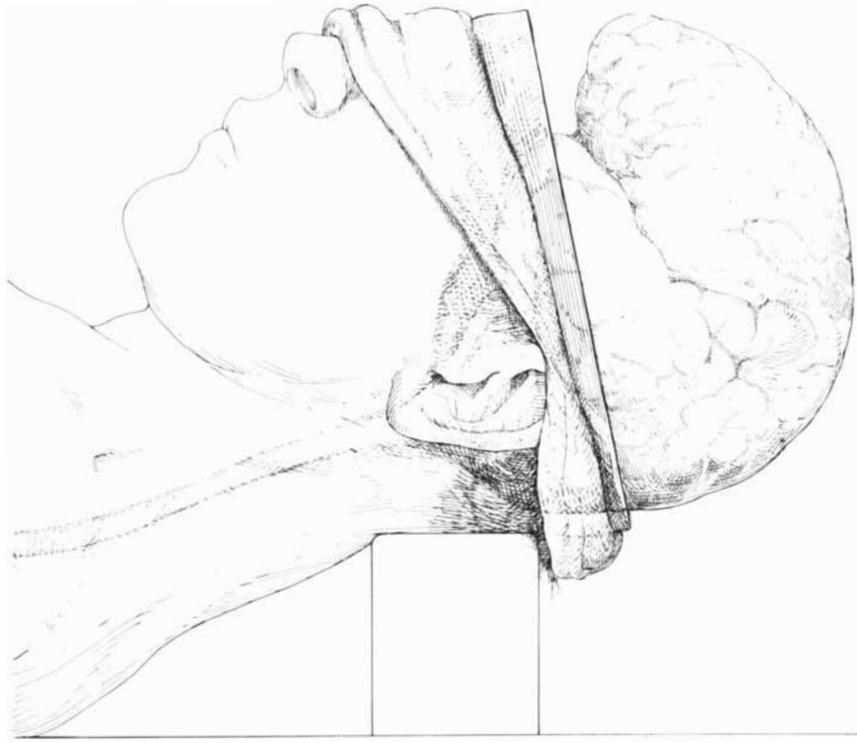


EVISCERATION OF ORGANS is done according to one of two different methods. Here the method is to remove the organs en masse and examine and dissect them as organ systems before dissecting the organs individually; the method was developed in the 19th century by the German physician Friedrich Albert von Zenker. In the other method, devised by Rudolf Virchow, the organs are removed from the body one by one and dissected separately. After the autopsy the organs are returned to the body and the incisions are sewn up.

ous hospitals. As is common practice in many institutions, each death certificate was filled out by one of the physicians responsible for the care, without benefit of the autopsy findings. Engel found that the underlying cause of death had been improperly or inaccurately recorded in 42 percent of the cases.

Since only 15 percent of the two million people who die in the U.S. each year are autopsied and the death certificates are often incorrect in even that small number of cases, it is reasonable to assume that each year at least a million death certificates are in error. How much of the nation's health-care planning is based on those certificates? Only postmortem examination can provide the basis for accurate statistics.

Similarly, only the systematic performance of autopsies can provide the data necessary for understanding the diseases that affect a nation's population. At present the pattern of autopsies is skewed. Of the 69 categories for cause of death listed in *Vital Statistics of the United States, 1977,* only six had autopsies performed in more than 50 percent of the cases: homicide, abortion, other complications of pregnancy, meningococcal infections, bacillary dysentery and amebiasis and "other external causes." These six conditions account for fewer than 2 percent of the deaths each year and are hardly the nation's most pressing health problems. In contrast only 9.9 percent of the more than 950,000 deaths due to major cardiovascular disease and 11.3 percent of the more than 385,000 deaths due to malignancy were autopsied.



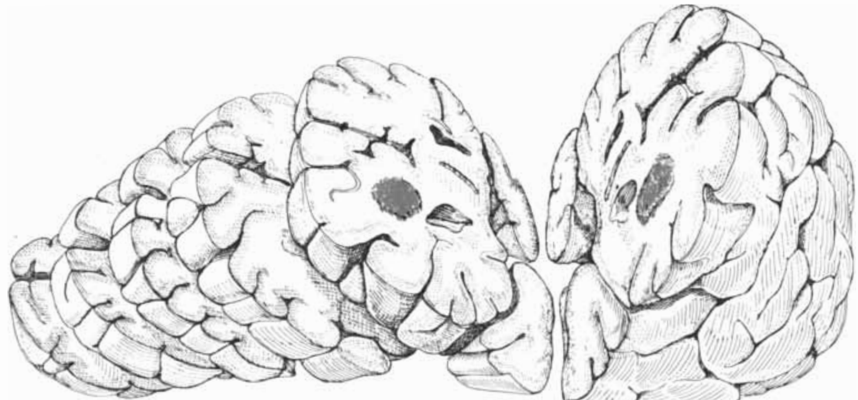
EXAMINATION OF THE BRAIN begins with a study of it in its normal position. The brain is then removed for closer examination. The spinal cord can also be removed through the skull.

Many other benefits of the autopsy can be cited. It is a vital complement to the considerable number of new diagnostic techniques that have been introduced during the past decade, such as computer-assisted tomography (the CAT scan), enabling the diagnostician to check the validity of the interpretations that have been made on the basis of readings from the machines. The autopsy is an excellent means of gauging the effectiveness of medications and, equally important, their relative toxicity, particularly for the many drugs employed in treating cancer. Those drugs often exert a strong effect on normal cells as well as on malignant ones. By applying the classic approach of observing the patient during life and correlating the findings at death with the clinical events the mechanisms of such toxic phenomena become known. For example, the tendency of Adriamycin, a potent anticancer agent, to cause heart failure became understood in this way, and the dosage was modified to minimize the likelihood of fatal injury.

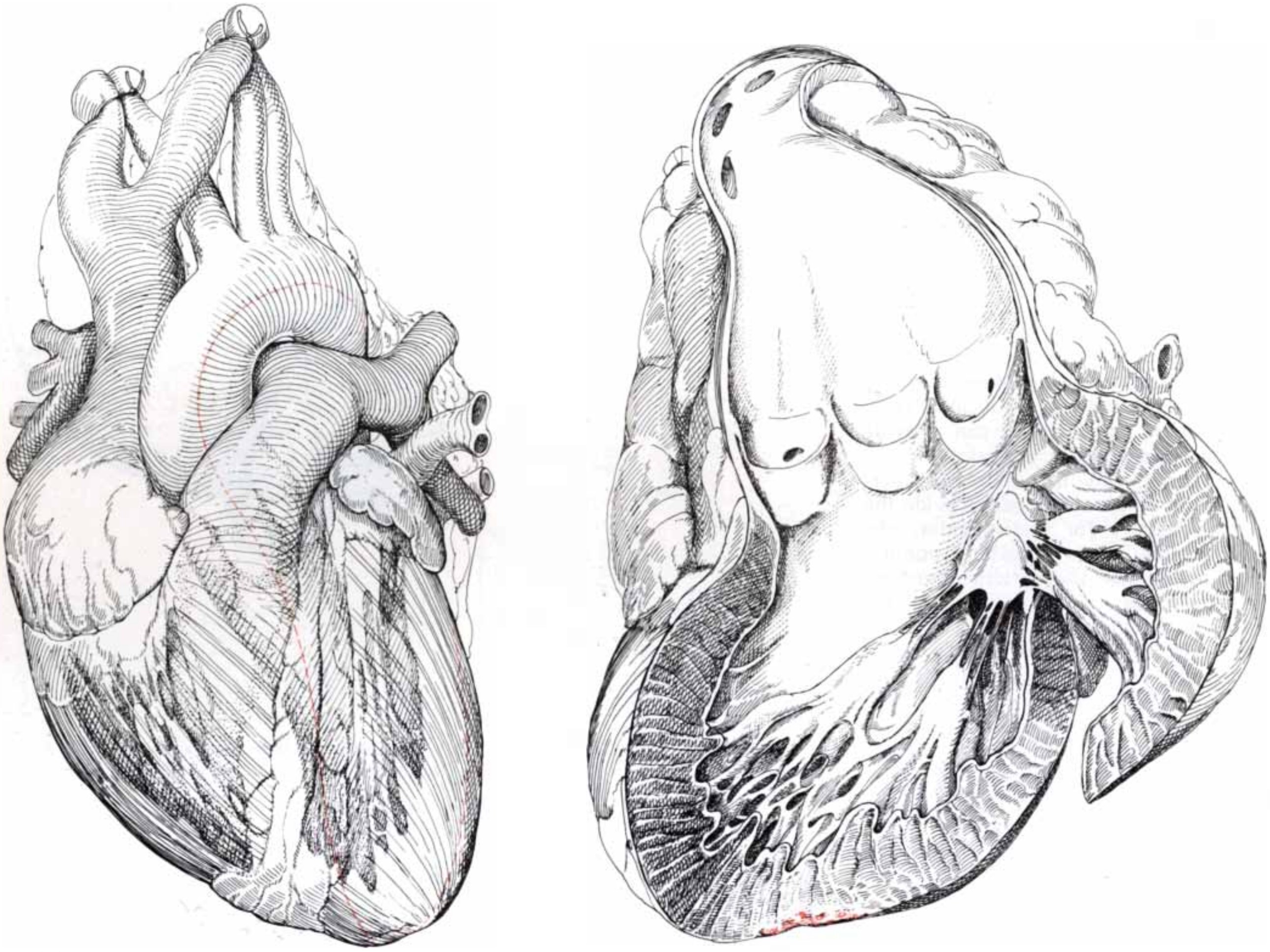
What are the long-range effects of prosthetic devices on tissues? In the early 1960's, when prosthetic heart valves were first being widely adopted, autopsies yielded the practical information on size, shape and materials that allowed for improvements in the design of the devices and for the refinement of the criteria employed to choose patients who were most likely to be helped by the artificial valves.

The autopsy is a singularly effective tool for recognizing new disorders that may be due to occupational or environmental influences. In the past decade autopsies helped to elucidate the basis of the development of a rare, highly malignant tumor—angiosarcoma of the liver—affecting workers in factories producing vinyl chloride. The repeated recognition of the tumor at autopsy was eventually followed by the appropriate epidemiologic and experimental studies, which led ultimately to the control of the conditions that gave rise to the tumor. The autopsy also helped to delineate the conditions associated with occupational exposure to asbestos.

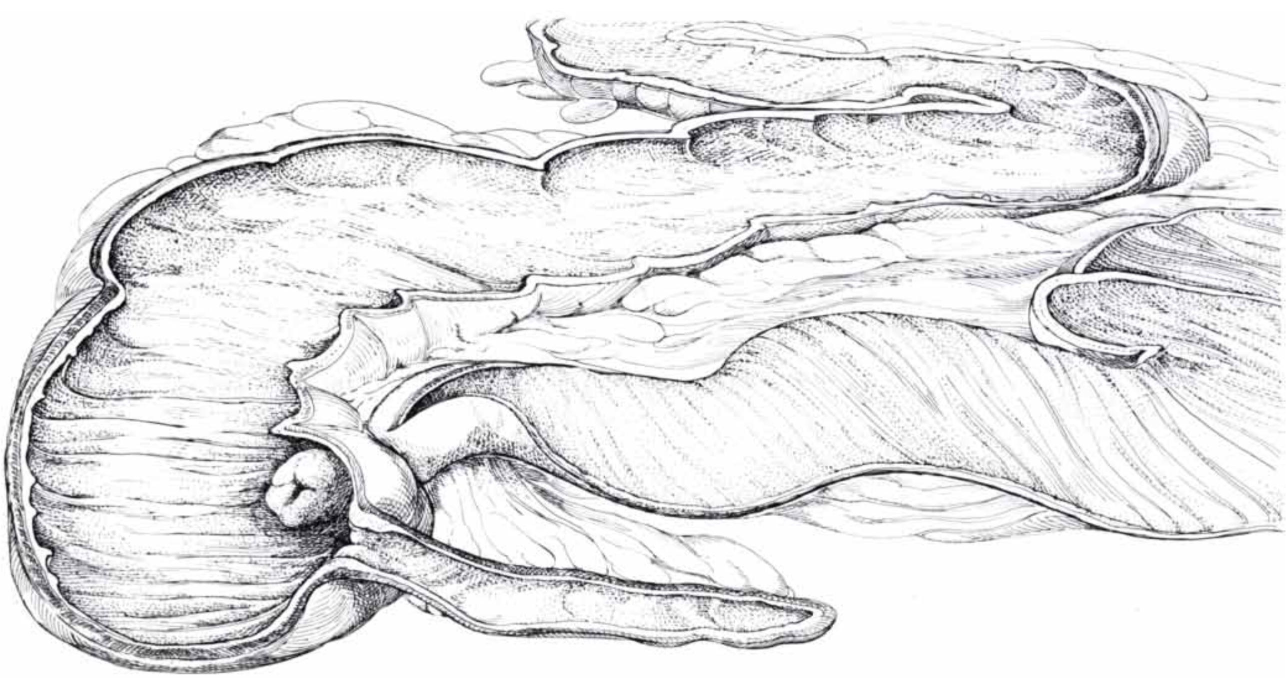
Environmental pathology is one of the newest, most publicized and least defined areas of study in medicine. The bulk of the evidence associating specific environmental factors with effects



NEXT STAGE in the examination of the brain is to make incisions in it in order to examine details and look for indications of disease. Here a lesion (*color*) is found; it is a tumor of the brain.



STUDY OF THE HEART, once it has been examined in relation to its associated blood vessels and the lungs, entails an incision (*broken line*) to open a ventricle, in this case the left one. When the ventricle has been opened (*right*), it is examined for evidence of disease. Here an infarct (*color*), an area where tissue has died when the circulation of blood was cut off by a clot, is seen in the bottom wall of the ventricle.



INTESTINE is examined separately after it has been studied in relation to the stomach and other associated organs, including the pancreas, the gall bladder and the bile duct. Here following dissection and an incision for closer inspection the intestine is found to be normal.

on organs remains circumstantial. The enormous potential of the autopsy to establish such relations is almost entirely unrealized. Some associations, such as air pollution and lung disease, are likely to be found. How many others can be identified and correlated with specific causative factors? What are the early, preclinical manifestations in the tissues of environmental injuries?

Organs and individual cells obtained at autopsy have been successfully maintained in culture for prolonged periods of time, making possible many in vitro experiments on human tissue. Human tissues recovered at autopsy can also be maintained in immunologically "nude" mice for many months, allowing for a multitude of in vivo experiments.

It is in the light of considerations such as these that one wonders why the rate of autopsy has declined so significantly over the past 30 years. A prominent reason is the feeling among many physicians that diagnoses are almost always established during life and that nothing new is learned from the autopsy. The statistics cited above on errors of diagnosis, together with the long list of diseases discovered or elucidated at autopsy in recent years (among them Legionnaires' disease and the toxic-shock syndrome), suggest that the opinion is insupportable.

A more important factor may be the reluctance of the physician to accept death, regarding it as a failure to be quickly forgotten. A compounding factor is that most patients who die in a hospital have been cared for by more than one physician. The trend toward specialization by physicians and the decrease in the number of general practitioners have led to a relative lack of familiarity between physician and patient. Doctors are no longer confidants and lifelong friends, and so they often lack the determination and confidence to pursue the matter of getting permission for an autopsy from the family at the time of death.

It is also possible that autopsy has declined because of a concern among physicians that it will disclose an error in diagnosis or treatment and an associated concern about malpractice suits. Although the data on this point are skimpy, documented instances in which the performance of an autopsy has been a major contributing factor in a malpractice suit are exceedingly rare. In fact, a thorough postmortem examination may support the defense of a physician whose medical care is shown by the examination to have been consistent with accepted practice.

Often the family of the person who has died is reluctant to give permission for an autopsy. The assertion "He's suffered enough" frequently goes unanswered, although the family member

Medicine is always changing, now there's a text that changes right along with it.

SCIENTIFIC AMERICAN *Medicine.*

The answers that you get from SCIENTIFIC AMERICAN *Medicine* are current, because each month new chapters replace the old. You will know where to look for the answers, because this advanced text follows the familiar structure of medical knowledge, changing as medicine changes.

Comprehensive. Leading authorities from Harvard and Stanford cover the full range of medicine for you the seasoned clinician.

Complimentary CME. At no extra cost, we offer a CME program in which you can earn 32 top credits* per year.

Trial Offer. We invite you to try SCIENTIFIC AMERICAN *Medicine* for two months with our compliments. Send us the coupon and you will receive the two-volume text and two monthly updates. You may also take a CME test for credit. At the end of 60 days, if you decide to continue the subscription, you will owe us $220 for the first 12 months (current renewal is $170); otherwise, return the books; there is no charge.

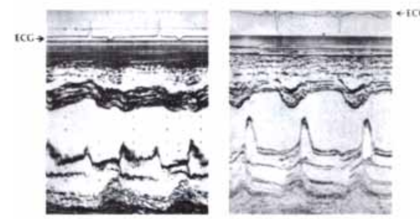
Please mail the coupon today so that we can put medical progress to work for you.

You may also order your trial copy by calling toll-free 1-800-345-8112 (in Pennsylvania, 1-800-662-2444).

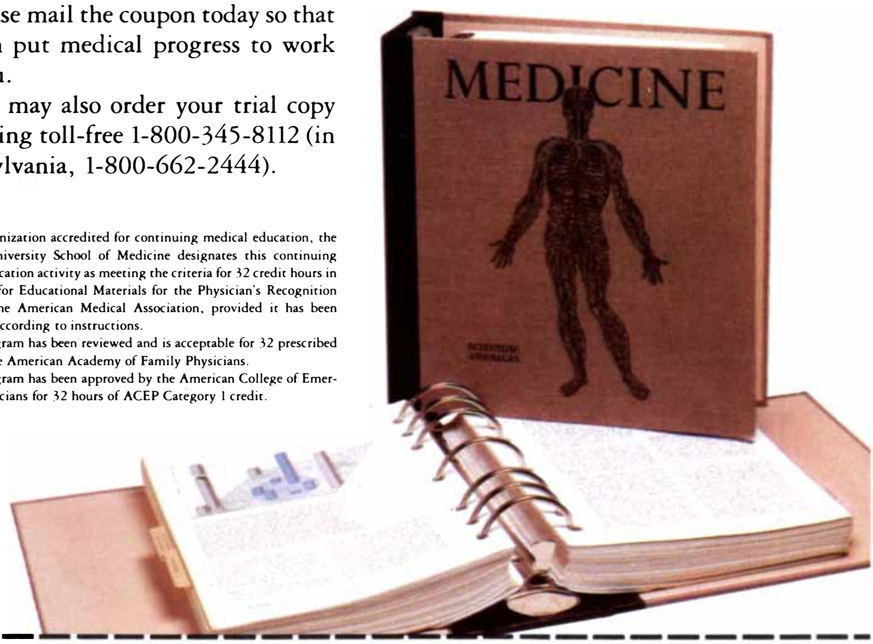
*As an organization accredited for continuing medical education, the Stanford University School of Medicine designates this continuing medical education activity as meeting the criteria for 32 credit hours in Category 1 for Educational Materials for the Physician's Recognition Award of the American Medical Association, provided it has been completed according to instructions.

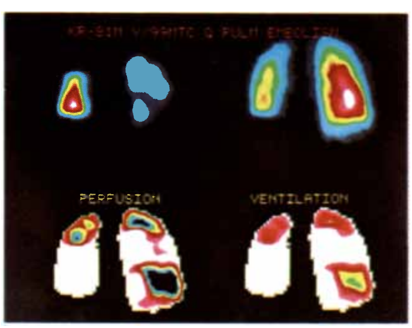
This program has been reviewed and is acceptable for 32 prescribed hours by the American Academy of Family Physicians.

This program has been approved by the American College of Emergency Physicians for 32 hours of ACEP Category 1 credit.

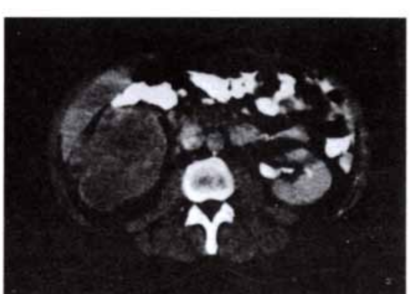


Echocardiograms from patients with aortic regurgitation.





Computerized scintigraphy reveals pulmonary thromboembolism.



Abdominal computed tomogram reveals large renal carcinoma replacing part of right kidney.

SCIENTIFIC AMERICAN MEDICINE

415 Madison Avenue, New York, New York 10017

Please enroll me as a subscriber to SCIENTIFIC AMERICAN *Medicine.* On receipt of this coupon you will send me the advanced two-volume text described in your announcement and update it regularly by sending me new monthly subsections. I understand that the annual subscription of $220 for SCIENTIFIC AMERICAN *Medicine* is tax deductible. If I am not entirely satisfied, I may cancel at any time during the first 60 days, returning all materials for a *complete refund.*

☐ Please enter my subscription for SCIENTIFIC AMERICAN *Medicine*
☐ I shall also enroll in the CME program
☐ I enclose a check made out to SCIENTIFIC AMERICAN *Medicine* for $220*
☐ Bill me ☐ VISA ☐ MasterCard

Expiration Date________ Account Number________

*Please add sales tax for California, Illinois, Massachusetts, Michigan and New York

Name________ MD Specialty________

Address________

City________ State________ Zip________

Signature________

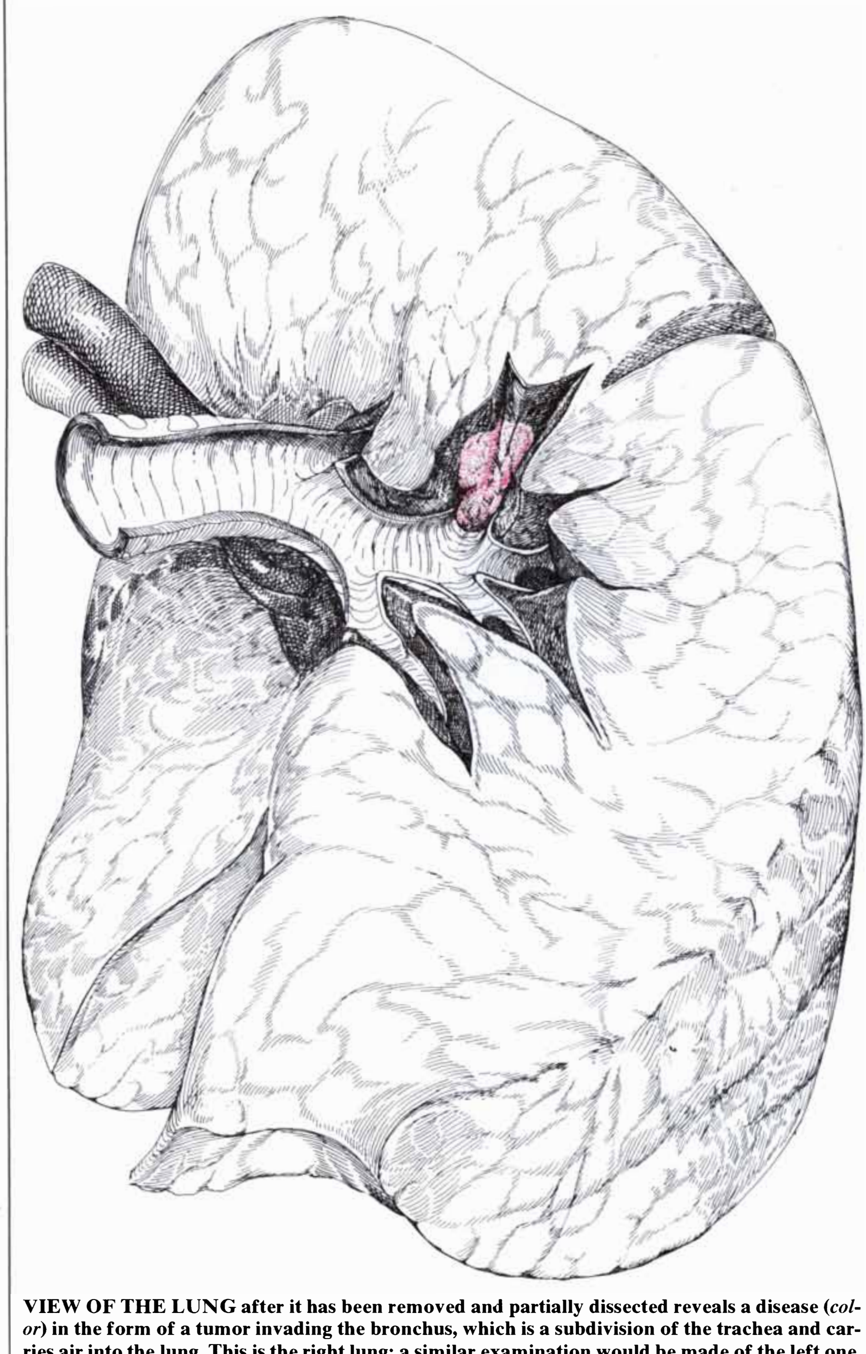
Please allow six to eight weeks for delivery. All payments must be in U.S. dollars. CME is available outside the U.S. and its possessions for an additional $65. Surface routes used unless airmail delivery is prepaid

7A

may actually be asking for reassurance that any suffering is over. Some families fear mutilation of the body, although this objection can usually be overcome by a simple description of the procedures. A family that has given permission for an autopsy in the past may not have been told of the results and so may see little reason for acceding again. Religious objections may be raised, although there are no edicts against autopsy by the Catholic church and the objections raised by orthodox Jews are based on rabbinical interpretations of the 18th century. In some communities undertakers discourage families from giving permission for autopsy because they do not want their own work delayed, even though the autopsy takes only a few hours.

Another factor in the decline of autopsy is that physicians have not had occasion to become interested in the subject. Pathology in general and autopsy in particular have been relatively deemphasized in medical schools. Participation in at least one autopsy was once a requirement for every medical student, but in many medical schools today it is not even an option.

In many hospitals the autopsy room is far removed from the areas where the living patients are. Clinicians think they do not have time to attend the autopsy,



VIEW OF THE LUNG after it has been removed and partially dissected reveals a disease (*color*) in the form of a tumor invading the bronchus, which is a subdivision of the trachea and carries air into the lung. This is the right lung; a similar examination would be made of the left one.

Osborne™ brings you the comparison IBM® and Apple® don't want you to see.

Other computer companies dazzle buyers with an array of options and add-ons that makes the final price hard to determine and makes the computer hard to buy, complex to assemble, and very difficult to carry.

We believe in making personal computers that are easy to learn and use. And that starts with making computers easy to *buy*.

The Osborne 1™ Personal Business Computer. One simple price, $1795, buys it all.

And it all comes in a portable case you can take with you wherever you work. Because once you go to work with an Osborne, you won't want to work any other way.

For your nearest dealer, call (in California) 800 772-3545, ext. 905; (outside California) call 800 227-1617, ext. 905.

OSBORNE™
COMPUTER CORPORATION

	OSBORNE 1™	IBM PERSONAL®	APPLE II®
Computer with 64K RAM, two floppy drives[A], keyboard and CRT:	$1795	$3240[B]	$3120[C]
Serial communications:	INCLUDED	EXTRA COST	EXTRA COST
Modem Connection:	INCLUDED	EXTRA COST	EXTRA COST
IEEE 488 Instrument communications:	INCLUDED	EXTRA COST	EXTRA COST
BASIC interpreter[D]:	INCLUDED	INCLUDED	INCLUDED
Business BASIC[E]:	INCLUDED	EXTRA COST	EXTRA COST
CP/M® Control Program:	INCLUDED	EXTRA COST	F (see below)
Word Processing[G]:	INCLUDED	EXTRA COST	EXTRA COST
Electronic Spreadsheet[H]:	INCLUDED	EXTRA COST	EXTRA COST
Carrying Case:	INCLUDED	EXTRA COST	EXTRA COST
TOTAL PRICE[I]:	**$1795**	**$4000–4700**	**$4000–4700**

A. The Osborne 1™ includes two built-in 100K byte floppy disk drives. The IBM® and APPLE II® drives provide approximately 160K bytes of storage. **B.** From the IBM Product Center Personal Computer Price Schedule. **C.** From the Apple Computer Suggested Retail Price List. **D.** The Osborne includes MBASIC® from Microsoft. **E.** The Osborne includes CBASIC®, a business-oriented BASIC language from Digital Research.™ **F.** The Osborne includes CP/M®, the industry-standard control program from Digital Research. The list of software packages which will run with CP/M is considerable. IBM offers CP/M 86 (a version of CP/M) at extra cost. There are optional hardware systems which allow the Apple II to run CP/M; the Apple II control program is highly comparable to CP/M. **G.** The Osborne includes WORDSTAR® word processing with MAILMERGE®—products of MicroPro™ International. **H.** The Osborne includes SUPERCALC™, the electronic spreadsheet system from Sorcim Corporation. **I.** Exact price comparisons cannot be presented, because the software and hardware options chosen to create the "equivalent" of the Osborne 1 Personal Business Computer vary in price. The range indicated was computed using price lists from IBM and Apple. Documentation of the computations are available on request from Osborne Computer Corporation. Trademarks: OSBORNE 1: Osborne Computer Corporation; SUPERCALC: Sorcim Corporation; Digital Research: Digital Research, Inc.; Registered Trademarks: WORDSTAR, MAILMERGE: MicroPro International Corporation of San Rafael, CA; MBASIC: Microsoft; CBASIC, CP/M: Digital Research, Inc.; IBM: IBM Corporation; Apple, Apple II: Apple Computer Corporation.


IT CAN ACTUALLY MAKE 55 M.P.H. INTERESTING.
If you think that's impossible, you've never driven the Volvo Turbo.
A car whose handling equipment can turn a curve in the road or a trip to the supermarket into a driving adventure.
Its turbocharged 4-cylinder engine can dust a V-8 off the line. Automotive writers have described it as "A blast." "Spectacular." "Like cutting in an afterburner."
Maybe you think speed limits, emissions controls and government mileage requirements have made driving humdrum. But that's only because you don't own an interesting car.
THE TURBO
By Volvo.

which often coincides with other activities in the hospital or with office hours. Even when an autopsy has been done, the results may not be communicated clearly and quickly to the responsible doctor. As William C. Roberts, chief of pathology at the National Heart, Lung, and Blood Institute, has pointed out, the pathology department in a hospital has many responsibilities, virtually all of them viewed as being more pressing than autopsies, so that "the autopsy reports are typed last by the secretaries, the paraffin blocks on autopsies are cut last by the technicians and the staff pathologists delay 'the signing out' of autopsies to the last because of more important duties." This almost inevitable delay further discourages physicians from seeking permission for autopsies.

Disenchantment also follows because of the disparity between the skills of the clinician and those of the person performing the autopsy, which is generally relegated to the most junior member of the pathology department. That person lacks the knowledge, judgment and experience to effectively translate the morphologic observations into clinically relevant comments. The autopsy is second to surgical pathology and the clinical laboratory, both of which serve the living directly, and these are the places where the senior members of the staff are most active.

Over the past 20 years cell biology has become more and more a part of the pathologist's responsibility and interest, the autopsy less and less. Senior pathologists have turned increasingly to sophisticated scientific or diagnostic endeavors. In a teaching institution the autopsy cannot compete with the exciting areas of surgical or experimental pathology as a basis for professional recognition and advancement. Senior pathologists seek other areas of academic expression and so become unavailable for immediate consultations with clinicians who need autopsy interpretations. The junior pathologist is deprived of any opportunity to learn from an instructive two-way discussion of clinical problems and concludes that the autopsy room is not the place to find success.

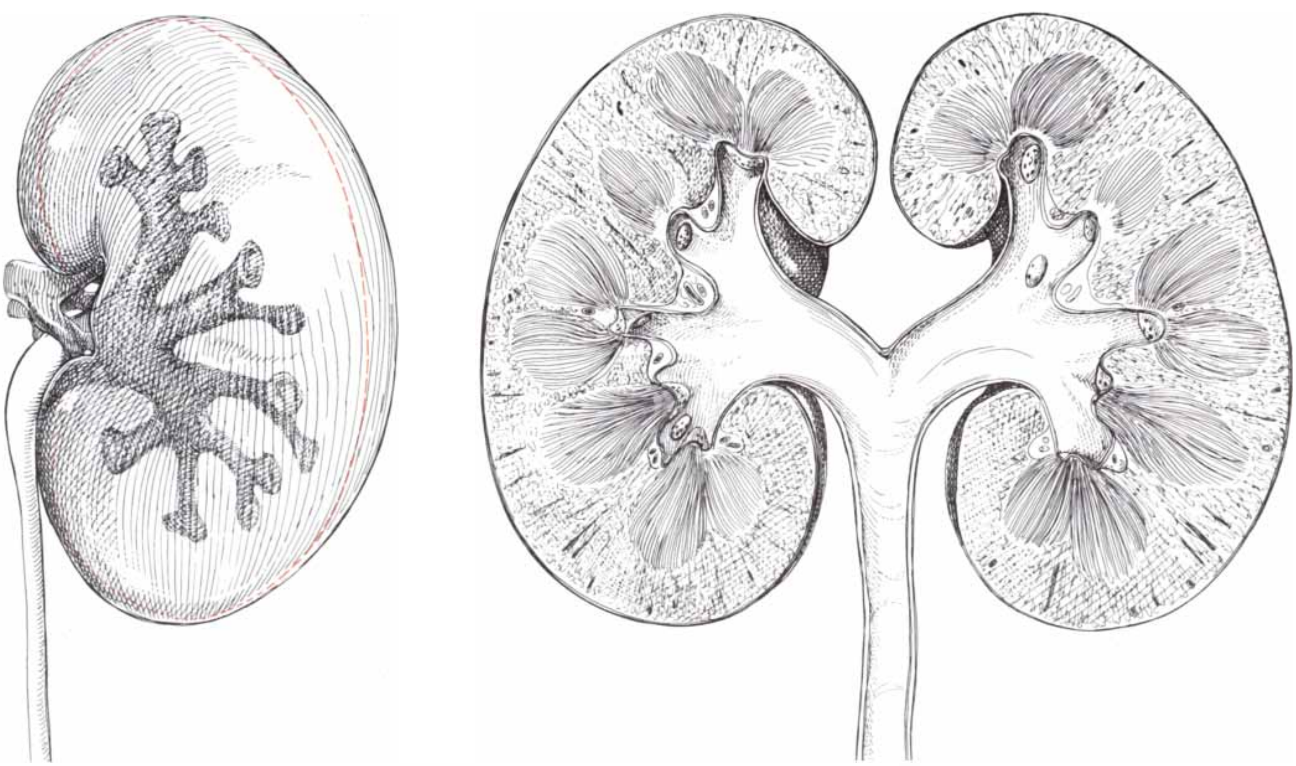
This unsatisfactory situation is further compounded when the autopsy is done by an autopsy technician or a pathologist's assistant rather than by a physician. Assistants of this kind have been utilized recently as a way of releasing the pathologist for ostensibly more important activities. The practice is also thought to have financial advantages. The trend may be shortsighted. If autopsy is to remain a significant component of the practice of medicine, it will require more of the pathologist's time and interest rather than less.

What can be done to reverse the decline of autopsy? To a great extent the pathologist is the key. The pathologist and the clinician must gain a renewed understanding of the critical role autopsy can play in health care. Hospital administrators must be made to understand the major service autopsy can provide for evaluating the cost-effectiveness of hospital activities and for maintaining excellence in medical care.

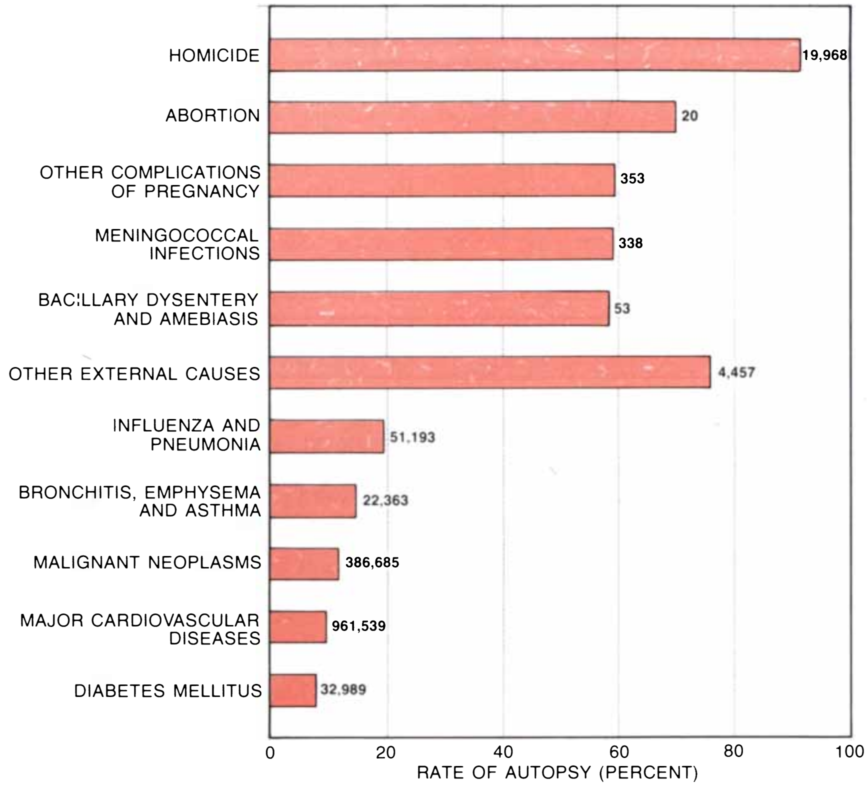
New approaches to the information-handling aspects of the autopsy would be helpful, but they cannot provide a remedy for the pathologist's failure to exercise a dynamic role. Before an autopsy is begun the pathologist and the clinician must identify together the major clinical problems that arose with the patient and the forms of therapy provided. After the autopsy the morphologic findings must be immediately presented to the physician to maximize their instructional value.

Families must be told of the results of the autopsy, and the significance of the key findings must be explained to them. A copy of the final report on the autopsy, with appropriate interpretations, should be given to the closest relative of the deceased. In addition the importance of autopsies should be more fully explained to the public. Hospitals should include a brief discussion of the advantages of the autopsy in the information booklets they give patients, just as they describe X rays, blood tests and many of their other services. Physicians as well as the public need to be told about the many recent contributions of autopsy to medical knowledge and about its vital role in monitoring the quality of medical care.

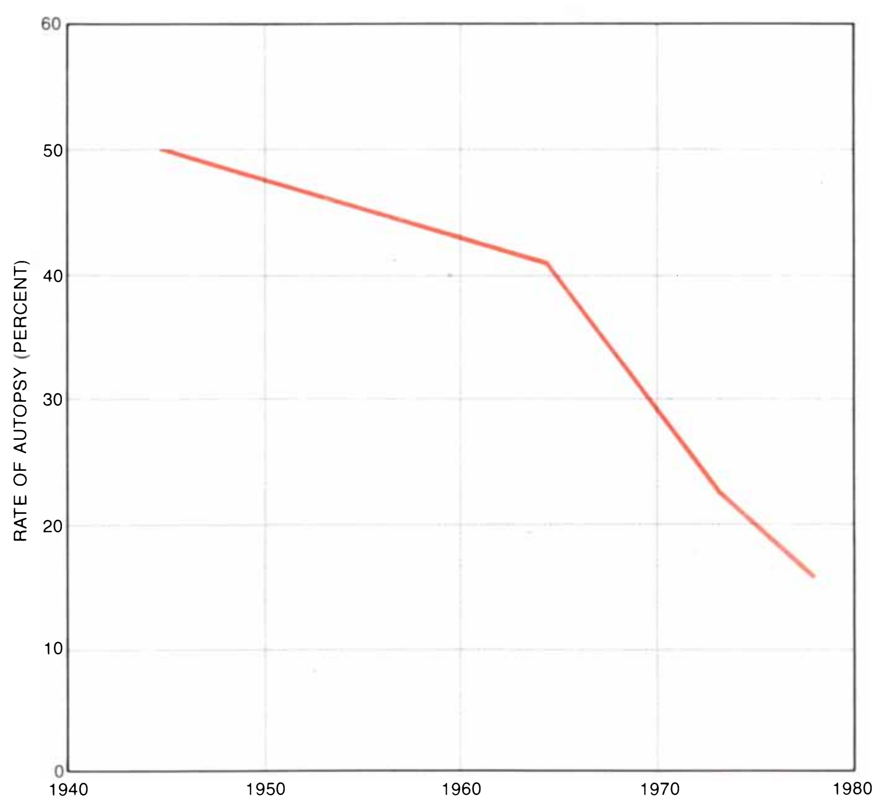
Senior pathologists must again de-



LEFT KIDNEY is shown (*left*) after it has been separately dissected from the organs associated with it. An incision has been made, as is indicated by the broken line, to open the kidney for an examination of its interior. No disease is found. The descending tube is the ureter.



SKEWED PATTERN OF AUTOPSIES is evident in a comparison of the six categories for cause of death in which the autopsy rate was 50 percent or higher in 1977 with five categories that accounted for far more deaths and included diseases that are among the nation's major health problems. The figure at the end of a bar shows the number of deaths in that category.



DECLINING RATE OF AUTOPSY in the U.S. is charted as a percentage of the number of deaths. One reason for the decline is a belief among physicians that little new is learned at autopsy, yet several diseases have recently been found or elucidated by the procedure.

velop and demonstrate skills in the performance of autopsies and must be able to solve clinical problems as well as morphologic ones. It used to be that physicians embarking on a career in pathology had at least one year of clinical experience. Similarly, clinicians spent some part of their training in a pathology department, often doing autopsies. This interchange of experience was thought to be particularly valuable for the physician in training and to be ultimately valuable for the patient. Pathologists were thought to gain a greater appreciation of the relevance of their activities, together with an ability to relate understandingly to their clinical colleagues. A step or two backward might be an advance.

An autopsy is a significant expense for a hospital, costing roughly $1,000, and it does not produce any income for either the hospital or the pathologist. The autopsy should be adequately funded so that it is not always second or lower in the line of priorities. Health-insurance agencies could benefit by providing direct reimbursement for the autopsy, since it yields many clues about disease and the causes of disease that can ultimately help to control the rising costs of health care.

At one time hospitals had to demonstrate an autopsy rate of 20 percent in order to be accredited. The requirement was eliminated in the 1960's. Data are lacking on what would be the most cost-effective and beneficial autopsy rate. Certainly it would not be worthwhile to do an autopsy in every case of death. William Roberts has suggested that some statistically determined sample of all people dying in a hospital should be autopsied rather than only those thought to present diagnostic problems, in order to achieve a reliable means of quality control. The ideal autopsy rate for a tertiary-care, teaching hospital is probably 100 percent. It is clearly not easy to determine the lowest acceptable rate.

John R. Carter, former director of the Institute of Pathology at Case Western Reserve University, is leading an effort by the College of American Pathologists to develop a computerized National Autopsy Data Bank, which would provide a central repository of pathological, biomedical, demographic and epidemiologic information. By this means the vast fund of medically and socially useful information yielded by autopsies would be readily accessible to everyone who might benefit from it. The data bank would surely bring about a far wider application of the inscription in the amphitheater where dissections were performed at the University of Bologna: *Hic locus est ubi mors gaudet succurere vitae* ("This place is where death rejoices to come to the aid of life").

Your future is not in the stars. But it may be in outer space.

Four out of five communications satellites in orbit communicate by means of traveling-wave tubes from Hughes Electron Dynamics.

And two out of five engineers in electronics have experience that Hughes in Torrance, outside Los Angeles, can use.

You can build something important here. State-of-the-art microwave and millimeter-wave components and systems. And a career. We are providing excellent salaries, outstanding benefits, and job stability for:

- **Microwave Engineers** — You need knowledge of electromagnetic theory in addition to your BS degree in electrical engineering, physics, or electronics.
- **Mechanical Engineers** — Can you perform mechanical design and analysis, in relation to the electrical and thermal requirements of spaceborne transmitters? If so, join us.
- **Technicians and Research Assistants** — If you have an AA degree or military or technical schooling — and an interest in RF, electricity, and mechanical matters — you can work with some of the best engineers and physicists in the world, on some of the most advanced communications projects in the world. And in space.
- **Reliability Engineers** — You will manage and perform reliability program tasks including: preparation of reliability plans, proposal inputs, failure analysis investigations, reliability demonstration testing, and considerable customer interface.

The stability of jobs here comes from our continuing growth and from the fact that Hughes Aircraft has more than 1,500 diversified projects underway.

And southern California is still a fine place to live and work. We don't see anybody leaving for the snow country, except for a weekend of skiing nearby.

We won't say "Put your career in orbit." We'll just say "Look into the Hughes Story. And write yourself in."

Send your resume to: Gary Adams, Hughes Electron Dynamics, Dept. AS-3, P.O. Box 2999, Torrance, CA 90509.

Proof of U.S. Citizenship Required
Equal Opportunity Employer

HUGHES

HUGHES AIRCRAFT COMPANY
ELECTRON DYNAMICS

THE AMATEUR SCIENTIST

How to analyze a city traffic-light system from the outside looking in

by Jearl Walker

If you drive an automobile in a city, you are sure to encounter a time when you have to stop for a red light at almost every intersection. Your progress is slow, your time is wasted and you may find yourself in a massive traffic jam. You think if only the sequence of traffic signals along the route were synchronized, you might be able to go indefinitely without stopping.

Actually in most cities the important traffic routes now have synchronized light systems, particularly during rush hours. When the drivers travel at a certain speed, they are supposed to find a green light at each intersection. In heavy traffic this sequencing forces the cars into what traffic analysts call platoons. In the gaps between them the traffic on perpendicular streets crosses the main street or merges onto it. On a properly designed route the sequence of lights moves the traffic efficiently without unduly delaying the perpendicular traffic.

To examine a synchronized light system I studied the lights along Carnegie Street in Cleveland. To be sure, I could have found out much of what I wanted to know by consulting city officials, but it would not have been as much fun. Besides, there is always the possibility that an independent analysis can come up with some useful ideas. You may want to try the same kind of analysis on your own local traffic system.

The lights along Carnegie Street are controlled by a preset electronic system rather than by a feedback system sensitive to the density of the traffic. The street is a major artery between downtown Cleveland and the eastern suburbs and is heavily traveled in the morning and afternoon rush hours. In the morning rush hour the westbound traffic has four lanes and the eastbound has two; in the afternoon rush hour the traffic is one-way eastbound. For the rest of the day the street has three-lane traffic in each direction.

Carnegie Street runs through an area that includes the Cleveland Play House and a variety of stores, homes and warehouses. Side streets, each three lanes wide, connect Carnegie with parallel streets that also carry rush-hour traffic. Since much traffic crosses between the major streets, the red lights on the side streets cannot stay on too long or the queues on them may grow to the point where they block the major streets.

I studied the system of traffic lights on Carnegie Street in the early stages of the afternoon rush hour. Beginning at the intersection of Carnegie and East 71st streets, I measured the distance from one intersection to the next as I walked along Carnegie in the direction of the traffic flow. At each intersection I timed the duration of the red, yellow and green lights. In addition as I proceeded I recorded the interval between the start of the green light at one intersection and the start of the green light at the next.

My tool for ascertaining times was a digital watch with a stopwatch feature and a particularly helpful split-time readout. I started the watch when the yellow light came on at an intersection. When the red or green light came on, I pushed the button for the split time. This function briefly held the display so that I could record the time in my notebook. The stopwatch was still running, however, and after about five seconds the display again recorded the passing seconds as before.

After many observations I averaged the readings for lights of each color in order to reduce the error arising from variations in the length of time it took me to react to a change of the light and to push the appropriate buttons on the watch. I also averaged my measurements of the time between the beginning of the green light at one intersection and that at the next intersection.

To measure the distances between intersections I counted paces between the stop lines at successive intersections. Later at home I walked at approximately the same pace across my yard. After measuring the distance with a meterstick I was able to convert my measurements of distance along Carnegie into units of meters. I walked the width of my yard rather than taking a single step in order to reduce the error that might result from variations in the length of my stride. Even so, I figure that the error in my measurements of the distance between the intersections is between 10 and 20 percent. Although surveying would be far more accurate, the added precision in my results would not be worth the extra effort.

The measurements are summarized in the illustration on the opposite page. The intersection of Carnegie and East 71st is represented at the lower left. The flow of traffic along Carnegie is upward in the drawing of the street. The other intersections are laid out along the drawing.

The horizontal scale of the graph represents the time of the lights. For example, the green light at Carnegie and East 71st is plotted as beginning at a time of five seconds. This is the bench mark from which that light and the ones at succeeding intersections were measured. I found that the green light at East 71st stays on for an average of 46.4 seconds. Hence the graph shows the onset of the yellow light at 51.4 seconds, the sum of 46.4 seconds and the bench mark of five seconds. The graph indicates similarly the start of the red light and the times for the next cycle of green, yellow and red.

The green light at Carnegie and East 77th, the next intersecting street, begins 29.5 seconds after the green begins at East 71st. Therefore the graph shows the start of the green phase at East 77th to be at 34.5 seconds, the sum of 29.5 seconds and the bench mark of five seconds. Since the green light there lasts for 48.6 seconds, the beginning of the yellow light is at 83.1 seconds on the graph. The color cycles for the other intersections are similarly plotted with respect to the bench mark for the green light at East 71st.

During the afternoon rush hour platoons of cars form on Carnegie near East 71st or somewhat closer to the downtown section. Usually a platoon moves eastward at a nearly constant speed. A platoon leader could race from intersection to intersection, but it would be to little advantage: over an extended time and distance one cannot go faster than the sequence of green lights allows. If the leader leaves one intersection at maximum acceleration and excessive speed, the car reaches the next intersection much too early for the green light.

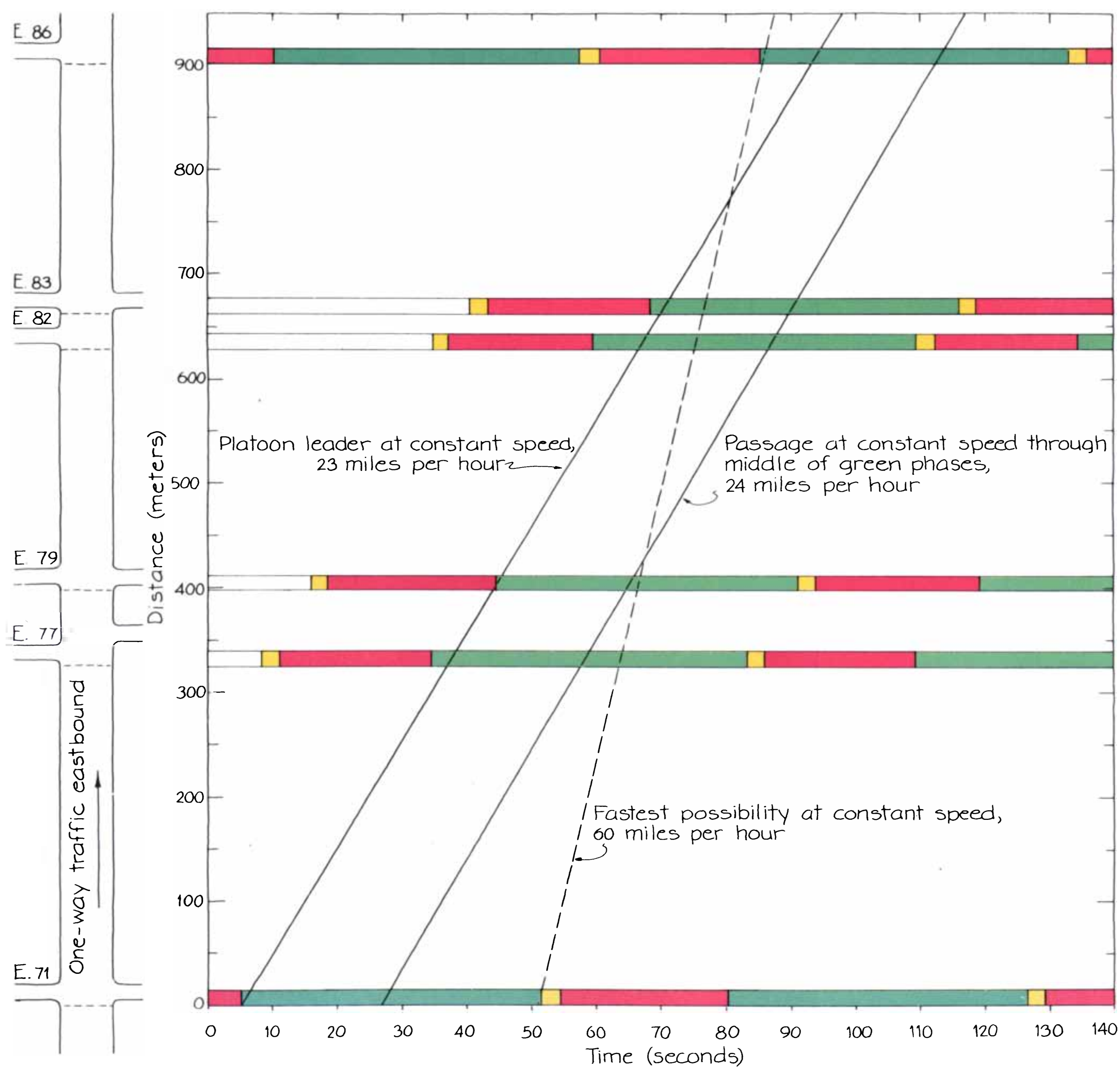
The graph indicates the motion of a platoon leader traveling at constant speed through the system of six intersections. The leader begins to move when the light at East 71st turns green and then travels along Carnegie at a constant speed to avoid red lights. According to my measurements, the light cycles are not synchronized perfectly for the leader. If the driver is to pass through the intersection at East 79th just at the onset of the green light, he reaches East 82nd about five seconds after the green light there has come on. Perhaps the mismatch is part of the design. It would allow the traffic merging onto Carnegie from East 79th to clear the intersections at East 82nd and East 83rd before the next platoon arrives.

The graph also indicates two other hypothetical drivers moving through the system at constant speed. One of them passes through each intersection at approximately the middle of the green phase. The other driver passes through East 71st precisely at the end of the green phase and then races to pass through East 86th just as the green light there comes on. The speed of the platoon leader and the other two drivers can be measured from the graph: the speed in meters per second is the slope of the line representing each driver.

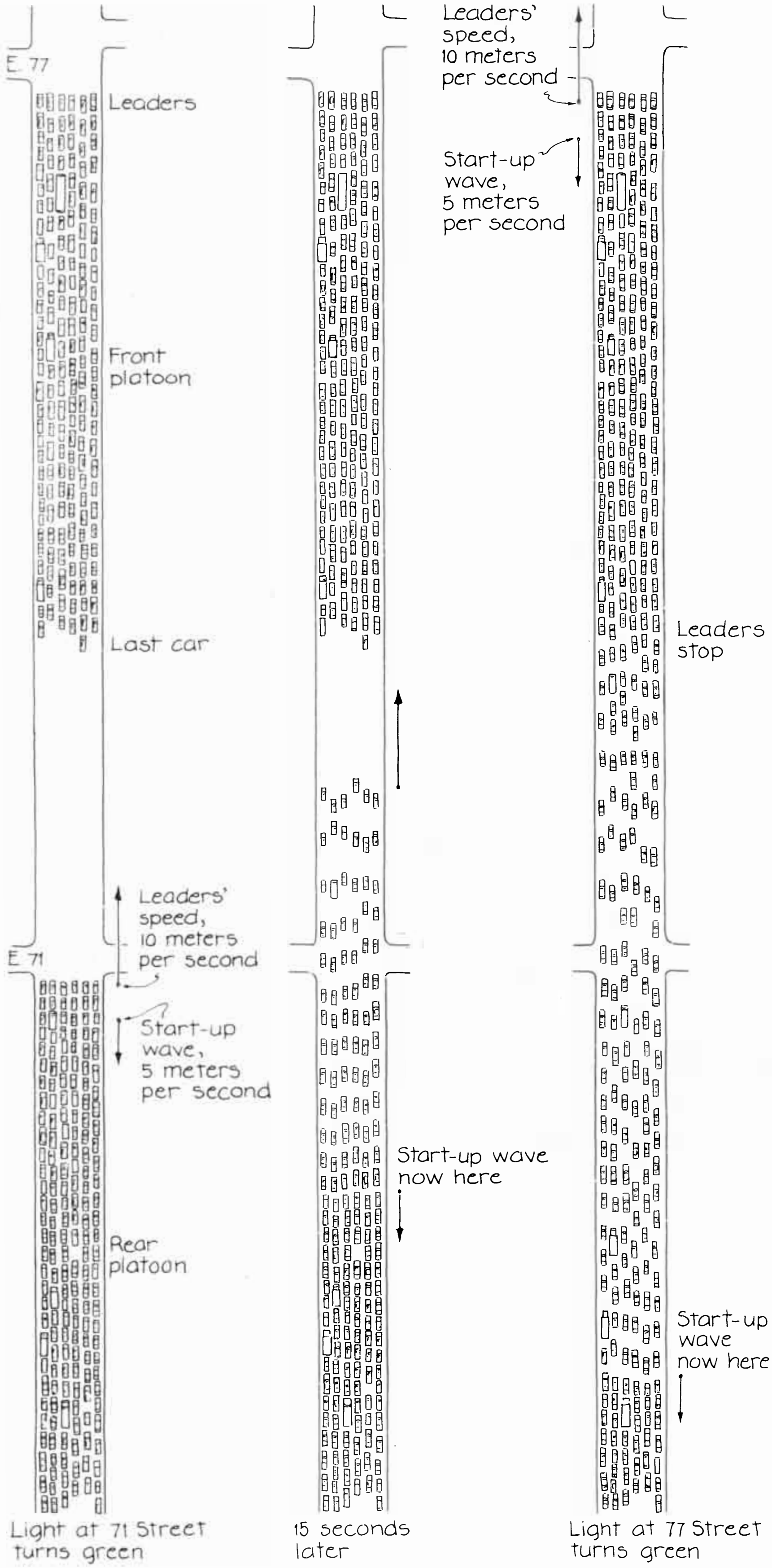
The platoon leader travels at 10.3 meters per second, which is about 23 miles per hour. The driver passing through each intersection at the middle of the green phase travels at 10.8 meters per second, which works out to approximately 24 m.p.h. The driver who races through the system in the least time moves at 26.6 meters per second, which is about 60 m.p.h. (The speed limit on Carnegie Street is 25 m.p.h.)

Compared with what happens in real traffic my law-abiding drivers are somewhat slow. When I drive through the rush-hour system on my way home every day, my platoon usually moves at about 28 m.p.h. Three factors contribute to the discrepancy between this speed and the speed computed from the graph. If the traffic flow is to be steady, a platoon leader must have a green light one or two seconds before arriving at an intersection, otherwise caution makes him slow down. The graph also ignores the time it takes the platoon leader to respond to the onset of the green light at East 71st. I figure that the reaction time is about a second.

The third factor is that the graph over-



The system of traffic signals on Carnegie Street in Cleveland during the afternoon rush hour



Conditions that can lead to a traffic jam

looks the time required to accelerate a car from rest to the constant speed that takes the platoon through the system. When I am the platoon leader, I need about three seconds to reach cruising speed. These three factors effectively reduce the time a platoon leader has to travel between intersections. Therefore he can drive at a constant speed of somewhat more than 23 m.p.h. without having to stop at any intersection.

At times I have driven through the system after Carnegie has become one-way but before the volume of traffic has built up. If I pass through East 71st near the end of the green phase, I can go faster than 50 m.p.h. through several of the intersections but must stop for the red light at the intersection past East 86th. Almost every time I have exceeded the speed limit I have been passed by another driver. Hence it is possible to travel through the system at a constant speed of about 60 m.p.h.

During a rush hour platoons of cars move along Carnegie about every 75 seconds, which is the interval between successive green lights at East 71st. Although the lights along the route are not synchronized perfectly, the system works well enough so that the platoons are not too large for the distance between successive intersections. In fact, the green phase for most of the intersections could be decreased by 10 or 15 seconds without delaying the average platoon. The green phases are apparently kept as long as they are in case something goes wrong. The extra time is needed if an accident clogs one of the lanes of traffic or the rush-hour traffic is notably heavy.

Most of the traffic on Carnegie in the early stage of the afternoon rush hour tends to stay toward the right side of the street, although many drivers avoid the rightmost lane because of the possibility of coming on a car that is disabled or illegally parked. As the volume of traffic builds up, the lanes become almost uniformly filled.

Until that time a platoon can normally maneuver around an obstacle in one of the lanes. When a lane of traffic slows, drivers at the rear of the queue pull into other lanes. An obstacle tends to force the traffic into a more uniform distribution among the lanes. Trouble develops when the other lanes are already full. Then the queue that forms behind an obstacle can disrupt the scheme designed to move platoons of cars through the system of lights.

The plan of the light system on Carnegie seems to assume that when a platoon leader approaches an intersection, the preceding platoon of cars will have moved on. If part of that platoon still blocks the way, the leader of the new one must slow down or stop. The entire timetable of lights can be thrown off.

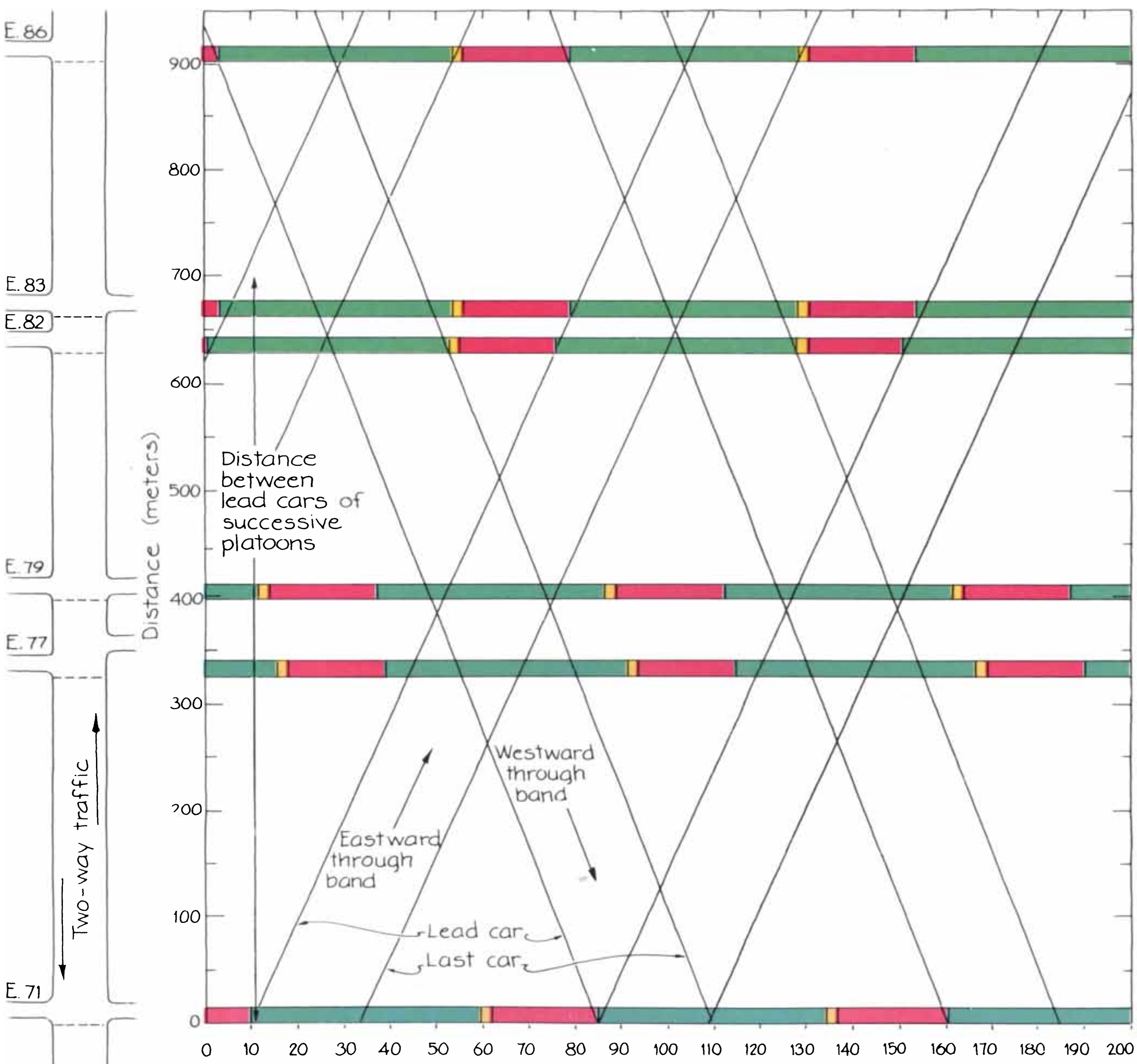
The delay may mean that part of the new platoon will not pass through the intersection before the next one arrives. If this happens, the problem becomes progressively worse.

The illustration on page 140 shows two platoons on Carnegie Street. The leading platoon is stopped at a red light at East 77th. The rear platoon has just got a green light at East 71st. First the leaders of the rear platoon begin to move into the intersection at East 71st. Then the cars behind the leaders move. This start-up motion travels the length of the platoon as a wave. Finally the last cars of the rear platoon move forward.

Once the leading platoon gets a green light it too will have a start-up wave propagating to its last cars. If they begin to move before the leaders of the rear platoon arrive, the light system works fine and all the cars move through it without unnecessary delay. Suppose, however, the rear platoon arrives too soon. Since its members must stop, the queue of cars gets longer and may eventually block the intersection at East 71st during the next change of lights there. The situation then has the makings of a traffic jam.

I have been caught in several traffic jams on Carnegie. Once it took me almost two hours to get through the six intersections from East 71st to East 86th. A heavy snowfall had delayed the early platoons of cars traveling through the system. Parts of the platoons were halted by red lights. Because of the slow driving on the snowy streets, the queues at the intersections lengthened as platoons from the rear arrived before the preceding platoons had moved on. The density of cars quickly became maximum: cars were bumper to bumper for miles along Carnegie. Since the intersections were blocked, the traffic on the perpendicular streets also became jammed. If traffic flow is considered to be an example of hydrodynamics, the traffic on that miserable winter afternoon was frozen solid.

For the situation depicted in the illustration on page 140 what length of the leading platoon causes the rear one to



The timing of signals for two-way traffic at midday be delayed? The critical length can be determined in terms of the distance between the intersections; it depends on the rate at which the start-up wave travels the length of the leading platoon. If the wave travels much slower than the speed at which the rear platoon approaches, even a short leading platoon can cause trouble.

To measure the speed of the start-up wave I stationed myself at intersections along Carnegie during a period of heavy traffic. I chose intersections that normally have queues of 10 cars or more. When the light for a queue turned green, I started my stopwatch. I stopped it when the last car in the queue began to move. I estimated the total distance between the front of the first car and the rear of the last. Dividing that distance by the time elapsed on the stopwatch, I estimated the speed of the wave. After many observations I found that the average speed was about five meters per second, or about half the designed cruise speed.

This finding means that the critical length for the leading platoon is only a small fraction of the total distance between East 71st and East 77th. If that platoon is more than about three car lengths long, the leaders of the rear platoon will be delayed. If both platoons are much longer than that, the rear platoon will not clear the intersection at East 71st before the next red light there.

Traffic jams along Carnegie might be avoided if the sequence of lights could be altered when the queues become too long for the normal rush-hour sequencing. The longer the leading platoon at East 77th is, the less head start the rear platoon should be given. If the length of the leading platoon is about a third of the distance between the two intersections, the green lights should come on simultaneously. If the leading platoon is even longer, the light at East 77th should come on before the one at East 71st, an order that is the reverse of the normal rush-hour sequencing. The head start given the leading platoon then allows the start-up wave to reach the last cars before the leaders of the rear platoon arrive.

In the bottom illustration on page 145 is an equation bearing on the platoon length at which the normal sequence must be changed. The length of the platoon (x) is expressed as a fraction of the distance between intersections (d). It depends on the speed of the start-up wave (v_1) and the normal cruise speed (v) for the system. When the platoons are shorter than the switchover value, the rear platoon should get the head start. When they are longer, the leading platoon should be given the head start. If the platoons are as long as the switchover value, they should get simultaneous green lights.

Some systems of traffic lights can be

changed in sequence when the queues become too long. The change could be preprogrammed into the system if long queues can be expected. In some systems the light sequence is affected by devices laid in the roadway to sense traffic flow. If Carnegie had such a system, my fellow prisoners and I might have been spared the ordeal of traveling only five blocks in two hours.

In the late morning and early afternoon Carnegie has three lanes of traffic in both directions. The sequencing of the lights should be different from what it is in the afternoon rush hour so that the westbound traffic will not be stopped too many times. I repeated my measurements of the light cycles at about noon. The results are given in the illustration on page 142, which shows the green light at East 71st starting at a time of 10 seconds. Thus 10 seconds is the bench mark for the color cycles of the intersections.

At each intersection the duration of the three colors of light is the same as it is during the rush hour, but the time between the onset of green lights at successive intersections is not the same. For example, during the rush hour the green light at East 79th comes on 15.4 seconds before the green light at East 82nd. The delay allows a platoon enough time to reach East 82nd before the green appears. For two-way traffic the system must be synchronized to minimize the delay for both directions of flow without forcing the perpendicular traffic to wait too long. At noon the green at East 82nd and East 83rd is already on when the green starts at East 79th. This system allows the westward traffic to travel uninterrupted through the intersections.

The illustration also shows "through bands," which are often put in such a graph of distance v. time. Here a through band indicates how a platoon traveling at a constant speed can pass through a system of intersections without pause. Although in light traffic cars could proceed at various speeds, the through bands are instructive to someone deciding how to synchronize the lights for most of the traffic. The illustration has two sets of through bands, one for eastward flow (the bands running upward to the right) and another for westward flow (the bands running downward to the right).

A through band consists of two parallel lines drawn as far apart as the graph for a traffic route allows. One of the eastward through bands in the illustration begins with the onset of the green light at East 71st. A straight line passes through the green phases of the other intersections up through East 86th. (The line is straight because it represents traffic moving at a constant speed.) Since the traffic is to flow along the route as quickly as it can without breaking the speed limit, the line on the graph should be as steep as possible. The line also must avoid any of the red phases because it represents uninterrupted traffic. Hence it skims across the end of the red phases at East 82nd and East 83rd.

The right side of the through band is parallel to the left side because it represents the rear cars of the platoon, all of whose members are traveling at the same speed. This line also is drawn to avoid any of the red phases. In the illustration the right side of the eastward through band skims the onset of red at East 86th.

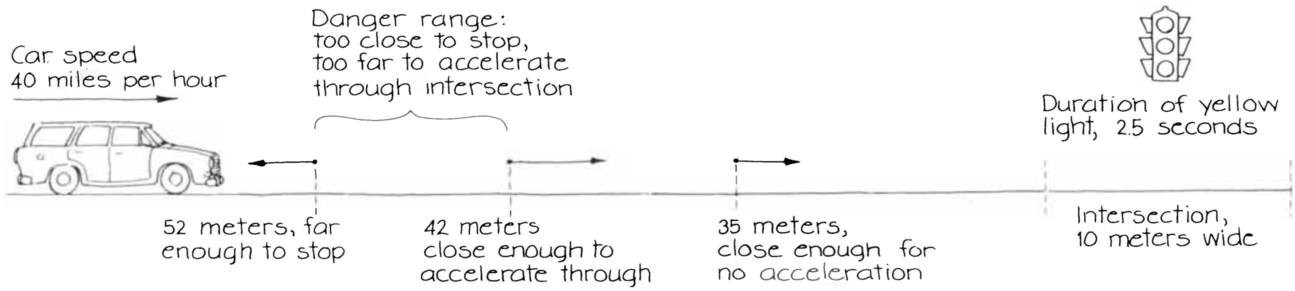
Suppose the eastward traffic along Carnegie at noon consists only of platoons of cars. The cruise speed of a platoon can be computed from the slope of the lines of the eastward through bands. If the traffic is to be uninterrupted, the platoon must begin at the onset of the green at East 71st and travel through the system of lights at a speed of 9.6 meters per second (equivalent to 21.5 m.p.h.).

How much later the rear of the platoon must pass through the intersection at East 71st is indicated by the right side of the through band. Apparently the rear must be no later than about 26 seconds after the green begins at that intersection. If a car in the platoon passes through the intersection later in the green phase there, it will be stopped by a red light at East 86th if not sooner. The through band during a rush hour is much wider, encompassing the entire green phase at East 71st. Thus when the Carnegie light system changes to the sequence for rush-hour traffic, the platoon speeds are not altered but the platoons can be much longer.

The distance between adjacent platoons can also be determined from the illustration. Start at the left side of one eastward through band, say at the bench mark for the onset of the green at East 71st. From the vertical axis read the distance to the left side of the through band between East 83rd and East 86th for that same time. The reading (about 700 meters) is the distance between the leading cars of the successive platoons. The distance between the last cars of the leading platoon and the first cars of the platoon beginning at East 71st is about 500 meters.

The illustration also includes information about the westward through bands. The left side of a band represents the travel of the first cars of a platoon. The right side represents the travel of the platoon's last cars. The westward and eastward through bands are approximately the same size, indicating that the system of lights is designed to pass traffic equally well in those directions at this time of day.

What is a suitable duration for the yellow light? It must be short enough to not



Strategies for a driver approaching a yellow light

delay the traffic unduly and long enough to allow a driver either to stop properly or to pass through the intersection before the red light comes on. The appropriate duration depends partly on the speed limit of the street, which determines a driver's ability to avoid being in the intersection when the red comes on. The higher the speed limit is, the longer the yellow light should be. This relation is sometimes ignored in setting the phases of a traffic light.

Assume that a yellow light lasts for 2.5 seconds, which is about the shortest duration I have found in Cleveland. Assume further that when the yellow first comes on, the driver immediately responds by braking to a stop. The minimum distance required for the stop depends on two factors: the car's initial speed (v_0) and the acceleration (a) provided by braking. (A physicist refers to any change in velocity as an acceleration; in common parlance the slowing of a car is called deceleration.) The stopping distance is computed as being the speed squared divided by twice the acceleration.

The acceleration of a braking car can vary greatly because it depends on the weather, the type of pavement, the condition of the tires and brakes and the skill of the driver. Usually the acceleration is initially between .9 meter and 3.1 meters per second per second. In the last stage of braking the acceleration is about 3.7 meters per second per second. In an emergency the acceleration could be as high as 6.1 meters per second per second, but braking at that rate is uncomfortable and unnerving.

Assume an acceleration of 3.1 meters per second per second. Suppose the driver is traveling at 40 m.p.h., which is 17.9 meters per second. The minimum stopping distance is 52 meters. If the driver is to stop before entering the intersection, the car must be no closer to the intersection than that distance. If it is closer, the driver has no reasonable choice except to continue through the intersection with no attempt at braking. Perhaps he should even accelerate in order to avoid being in the intersection when the red light comes on.

How successful can the driver be in such an acceleration? What he can do depends not only on the car's pickup but also on the initial distance of the car from the intersection. If the intersection is too far away, the driver cannot reach the far side before the red light comes on. The formula in the bottom illustration on this page yields the maximum distance the car must be from the intersection if the attempt to get through is to be successful. The formula encompasses the duration of the yellow light (t) and the acceleration (a) and initial speed (v_0) of the car. It also requires the width of the intersection (s) because the driver must reach the far side while the yellow light is still on.

A typical car can accelerate at from one meter to 2.2 meters per second per second. A sports car can probably accelerate at twice that rate. Consider a car with an acceleration of 2.2 meters per second per second. Assume that the driver responds instantly to the onset of the yellow light. Take the duration of the yellow light as being 2.5 seconds and the width of the intersection as being 10 meters. If the car is initially traveling at 40 m.p.h., the maximum distance between it and the intersection should be about 42 meters; if it is more than that, the car will be in the intersection during part of the red light.

The driver may decide to go through the yellow light without accelerating beyond the speed limit. If the initial speed of 40 m.p.h. is maintained, the car should be no farther than 35 meters from the intersection.

I have chosen the above numbers in order to make a point. If the driver is to stop successfully, the car must be no closer than 52 meters from the intersection. If he is to race through at maximum acceleration, the car must be no farther than 42 meters from the intersection. Between those limits what should he do? In principle neither strategy will succeed. The dangerous region between the two limits is larger with higher initial speeds for the car. Under the same assumptions as before except for an initial speed of 55 m.p.h., the danger region is 39 meters long. If the driver decides not to accelerate, the danger region is 46 meters long.

I once drove through such an intersection on a highway that had a speed limit of 55 m.p.h. I found myself facing a yellow light with neither the space to stop nor the acceleration to race through before the red light came on. I was saved from the possibility of a collision only by a delay in the light system: the green light for the perpendicular traffic came on about a second or so after the yellow light ended.

This month's bibliography [*next page*] cites two articles in which teachers of physics have analyzed the dilemma of a short yellow light. You might want to examine yellow lights in your neighborhood. Please keep in mind that the values I have used for the accelerations of a car are assumed. The braking acceleration is particularly open to question. It depends on the coefficient of friction between the tires and the street. If the street is covered with ice, snow, rain or anything else that reduces friction, the coefficient can be much lower than normal. The distance required for stopping the car is then much longer. Therefore the danger region is also longer and the possibility of a collision between a car and a vehicle in the perpendicular traffic is greater.

Switchover length for successive platoons:

$$\frac{x}{d} = \frac{v_1}{(v_1 + v)}$$

Minimum distance to brake a car:

$$d = \frac{v_0^2}{2a}$$

Maximum distance to race through a yellow light:

$$d = v_0 t + \frac{1}{2} a t^2 - s$$

Maximum distance to drive at constant speed through a yellow light:

$$d = v_0 t - s$$

Equations for traffic flow

BIBLIOGRAPHY

Readers interested in further explanation of the subjects covered by the articles in this issue may find the following lists of publications helpful.

METAMAGICAL THEMAS

A NEW UCI LISP MANUAL. Edited by John R. Meehan. Lawrence Erlbaum Associates, Inc., 1979.

A MICRO-MANUAL FOR LISP—NOT THE WHOLE TRUTH. John McCarthy in *History of Programming Languages,* edited by Richard L. Wexelblat. ACM Monograph Series, Academic Press, 1980.

THE LAW OF THE SEA

OCEAN YEARBOOK. Edited by Elisabeth Mann Borgese and Norton Ginzburg. International Ocean Institute, University of Chicago Press, 1979–82.

OCEAN MANAGEMENT: A NEW PERSPECTIVE. John M. Armstrong and Peter C. Ryner. Ann Arbor Science Publishers/The Butterworth Group, 1981.

MICROPROGRAMMING

AN OVERVIEW OF FIRMWARE ENGINEERING. Scott Davidson and Bruce Shriver in *Computer,* Vol. 11, No. 5, pages 21–33; May, 1978.

MICROPROGRAMMED CONTROL AND RELIABLE DESIGN OF SMALL COMPUTERS. George D. Kraft and Wing N. Toy. Prentice-Hall, Inc., 1981.

MORE HARDWARE MEANS LESS SOFTWARE. Robert Bernhard in *IEEE Spectrum,* Vol. 15, No. 12, pages 30–37; December, 1981.

THE SCRIPT OF THE INDUS VALLEY CIVILIZATION

THE SCRIPT OF HARAPPA AND MOHENJODARO AND ITS CONNECTION WITH OTHER SCRIPTS. G. R. Hunter. Kegan Paul, Trench, Trubner & Co., Ltd., 1934.

A DRAVIDIAN ETYMOLOGICAL DICTIONARY. T. Burrow and M. B. Emeneau. Oxford University Press, 1961.

THE INDUS CIVILIZATION. Sir Mortimer Wheeler. Cambridge University Press, 1968.

THE ROOTS OF ANCIENT INDIA. Walter A. Fairservis, Jr. The University of Chicago Press, 1975.

THE INDUS SCRIPTS—TEXTS, CONCORDANCE AND TABLES. I. Mahadevan. Memoirs of the Archaeological Survey of India No. 77, Manager of Publications, Government of India, 1977.

A SKETCH OF COMPARATIVE DRAVIDIAN MORPHOLOGY. Kamil V. Zvelebil. Mouton Publishers, The Hague, 1978.

MITOCHONDRIAL DNA

SEQUENCE OF INTRONS AND FLANKING EXONS IN WILD-TYPE AND *BOX 3* MUTANTS OF CYTOCHROME B REVEALS AN INTERLACED SPLICING PROTEIN CODED BY AN INTRON. Jaga Lazowska, Claude Jacq and Piotr P. Slonimski in *Cell,* Vol. 22, No. 2, Part 2, pages 333–348; November, 1980.

MITOCHONDRIAL GENES. Edited by P. P. Slonimski, P. Borst and G. Attardi. Cold Spring Harbor Monographs, Vol. 12, 1981.

SEQUENCE AND ORGANIZATION OF HUMAN MITOCHONDRIAL GENOME. S. Anderson, A. T. Bankier, B. G. Barrell, M. H. L. de Bruijn, A. R. Coulson, J. Drouin, I. C. Eperon, D. P. Nierlich, B. A. Roe, F. Sanger, P. H. Schreier, A. J. H. Smith, R. Staden and I. G. Young in *Nature,* Vol. 290, No. 5806, pages 457–465; April 9, 1981.

THE FUTURE OF THE UNIVERSE

THE FIRST THREE MINUTES: A MODERN VIEW OF THE ORIGIN OF THE UNIVERSE. Steven Weinberg. Basic Books, Inc., 1977.

TIME WITHOUT END: PHYSICS AND BIOLOGY IN AN OPEN UNIVERSE. Freeman J. Dyson in *Reviews of Modern Physics,* Vol. 51, No. 3, pages 447–460; July, 1979.

EFFECTS OF PROTON DECAY ON THE COSMOLOGICAL FUTURE. Duane A. Dicus, John R. Letaw, Doris C. Teplitz and Vigdor L. Teplitz in *The Astrophysical Journal,* Vol. 252, No. 1, pages 1–9; January 1, 1982.

MEMORY IN FOOD-HOARDING BIRDS

FOOD STORING BY MARSH TITS. R. J. Cowie, J. R. Krebs and D. F. Sherry in *Animal Behaviour,* Vol. 29, pages 1252–1259; 1981.

MEMORY FOR THE LOCATION OF STORED FOOD IN MARSH TITS. D. F. Sherry, J. R. Krebs and R. J. Cowie in *Animal Behaviour,* Vol. 29, pages 1260–1266; 1981.

AN EXPERIMENTAL ANALYSIS OF CACHE RECOVERY IN CLARK'S NUTCRACKER. S. B. Vander Wall in *Animal Behaviour,* Vol. 30, pages 84–94; 1982.

OSCILLATING CHEMICAL REACTIONS

OSCILLATIONS IN CHEMICAL SYSTEMS, II: THOROUGH ANALYSIS OF TEMPORAL OSCILLATION IN THE BROMATE-CERIUM-MALONIC ACID SYSTEM. Richard J. Field, Endre Körös and Richard M. Noyes in *Journal of the American Chemical Society,* Vol. 94, No. 25, pages 8649–8664; December 13, 1972.

BIOLOGICAL AND BIOCHEMICAL OSCILLATORS: PROCEEDINGS OF THE CONFERENCE ON BIOLOGICAL AND BIOCHEMICAL OSCILLATORS, PRAGUE, 1968. Edited by B. Chance, E. K. Pye, A. K. Ghosh and B. Hess. Academic Press, 1973.

CHEMICAL OSCILLATIONS. G. Nicolis and J. Portnow in *Chemical Reviews,* Vol. 73, No. 4, pages 365–384; August, 1973.

A SYSTEMATICALLY DESIGNED HOMOGENEOUS OSCILLATING REACTION: THE ARSENITE-IODATE-CHLORITE SYSTEM. Patrick De Kepper, Irving R. Epstein and Kenneth Kustin in *Journal of the American Chemical Society,* Vol. 103, No. 8, pages 2133–2134; April 22, 1981.

MINIMAL BROMATE OSCILLATOR: BROMATE-BROMIDE-CATALYST. Miklós Orbán, Patrick De Kepper and Irving R. Epstein in *Journal of the American Chemical Society,* Vol. 104, No. 9, pages 2657–2658; May 5, 1982.

AUTOPSY

A HISTORY OF THE AUTOPSY. Lester S. King and Marjorie C. Meehan in *American Journal of Pathology,* Vol. 73, No. 2, pages 514–544; November, 1973.

THE AUTOPSY: ITS DECLINE AND A SUGGESTION FOR ITS REVIVAL. William C. Roberts in *The New England Journal of Medicine,* Vol. 299, No. 7, pages 332–338; August 17, 1978.

TO SEE OR NOT TO SEE: THE STATUS OF THE AUTOPSY AT THE MOUNT SINAI MEDICAL CENTER. Stephen A. Geller in *The Mount Sinai Journal of Medicine,* Vol. 46, No. 1, pages 33–38; January–February, 1979.

PROPOSAL FOR A NATIONAL AUTOPSY DATA BANK: A POTENTIAL CONTRIBUTION OF PATHOLOGISTS TO THE HEALTH CARE OF THE NATION. John R. Carter, Nancy P. Nash, Ronald L. Cechner and Ron D. Platt in *American Journal of Clinical Pathology,* Vol. 74, No. 4 Supplement, pages 597–617; October, 1981.

THE AMATEUR SCIENTIST

THE STOP-LIGHT DILEMMA. Howard S. Seifert in *American Journal of Physics,* Vol. 30, pages 216–218; 1962.

TO STOP OR NOT TO STOP—KINEMATICS AND THE YELLOW LIGHT. J. Fred Watts in *The Physics Teacher,* Vol. 19, No. 2, pages 114–115; February, 1981.


1983 FORD LTD CROWN VICTORIA
IT'S A COMFORT TO KNOW YOU CAN STILL OWN THIS MUCH CAR.
Crown Victoria for 1983. Quiet. Smooth riding. Full size. With comfort and luxury for six passengers. Standard this year is a 5.0 liter Electronic Fuel Injected engine with automatic overdrive transmission. Rich velour fabrics, thick carpeting and warm woodtones create a quiet place for you to relax. While a solid steel frame and remarkable sound insulation make it clear you're riding in a truly fine automobile.
LTD Crown Victoria for 1983. In two- and four-door models. It's so rewarding to own one. Isn't it nice to know you still can?
Get it together — Buckle up.
HAVE YOU DRIVEN A FORD...LATELY?
Ford
FORD DIVISION


JAMESON
ESTD
1780
IRISH WHISKEY
A BLEND
DISTILLED IN IRELAND BY
LIMITED
DUBLIN
IMPORTED
Now that you're ready for a change of pace
it's time to try John Jameson.
Take a sip of John Jameson. Note the light, delicate taste.
Luxurious and smooth as you would expect a premium whiskey to be.
But with a distinctive character all its own. Set a new pace
for yourself. Step ahead of the crowd with John Jameson, the
world's largest selling Irish Whiskey.
80 PROOF CALVERT DISTILLERS CO., N.Y.C.